Semantic Segmentation without Annotating Segments

Wei Xia*, Csaba Domokos*, Jian Dong, Loong-Fah Cheong and Shuicheng Yan
Dept. of ECE, National University of Singapore, Singapore, 117576
{weixia,eledc,a0068947,eleclf,eleyans}@nus.edu.sg

Abstract

Numerous existing object segmentation frameworks commonly utilize the object bounding box as a prior. In this paper, we address semantic segmentation assuming that object bounding boxes are provided by object detectors, but no training data with annotated segments are available. Based on a set of segment hypotheses, we introduce a simple voting scheme to estimate shape guidance for each bounding box. The derived shape guidance is used in the subsequent graph-cut-based figure-ground segmentation. The final segmentation result is obtained by merging the segmentation results in the bounding boxes. We conduct an extensive analysis of the effect of object bounding box accuracy. Comprehensive experiments on both the challenging PASCAL VOC object segmentation dataset and GrabCut-50 image segmentation dataset show that the proposed approach achieves competitive results compared to previous detection or bounding box prior based methods, as well as other state-of-the-art semantic segmentation methods.

1. Introduction

Object classification, detection and segmentation are the core and strongly correlated sub-tasks [21, 28, 5] of object recognition, each yielding different levels of understanding. The classification tells *what objects* the image contains, detection further solves the problem of *where* the objects are in the image, while segmentation aims to *assign class label* to each pixel. In the case of *semantic segmentation* (see Fig. 1), the possible class labels are from a predefined set, which has attracted wide interest in computer vision [18, 19, 5, 1, 3, 8].

Bottom-up approaches extract various low and mid-level image features and try to find homogeneous segments based on these image cues. Li *et al*. [7] proposed a method where figure-ground hypotheses are generated by solving *constrained parametric min-cut* (CPMC) [8] problems with various choices of the parameter. The hypotheses are ranked and classified by making use of support vector regression (SVR) based on their "objectness". Analogous to average

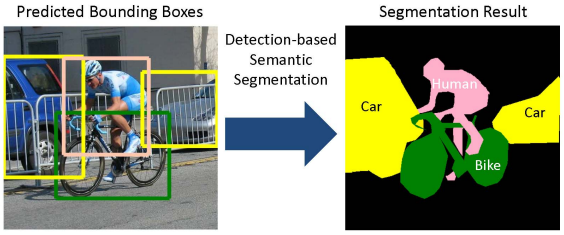


Figure 1. Semantic segmentation by using object bounding boxes.

and max-pooling, *second order pooling* is applied in [6] to encode the second order statistics of local descriptors inside a region. By applying this pooling technique a significant improvement can be achieved leading to the state-of-the-art performance [12]. CPMC-based works [7, 8, 6] alleviate the problem by exploiting object-level segments that have quite high overlap with ground truth objects. However, they still cannot guarantee the perfect classification and ranking of the segments, especially for visually confusing categories (*e.g*. cats and dogs).

Küttel *et al*. [17] proposed a *figure-ground segmentation* framework, in which the training masks are transferred to object windows on the test image based on visual similarity. Then, these masks are used to derive appearance and location information for graph-cut-based minimization. In [15], similar idea is proposed and a class-independent shape prior is introduced to transfer object shapes from an exemplar database to the test image. This prior information is enforced in a graph-cut formulation to obtain figure-ground segmentation. Generally, bottom-up methods without modelling objects globally tend to generate visually consistent segmentation instead of semantically meaningful ones.

Top-down approaches generally rely on acquired class-specific information. Shape model can also guide top-down segmentation. Brox *et al*. [5] applied so-called *poselets* to predict masks for numerous parts of an object. The poselets are aligned to the object contours, and then they are aggregated into an object. Arbeláez *et al*. [1] proposed region-based object detectors that integrate top-down poselet detector and global appearance cues. This method [1] produces class-specific scores for the regions and aggregates multiple overlapping candidates through pixel classification

*Wei Xia and Csaba Domokos contributed equally to this paper.

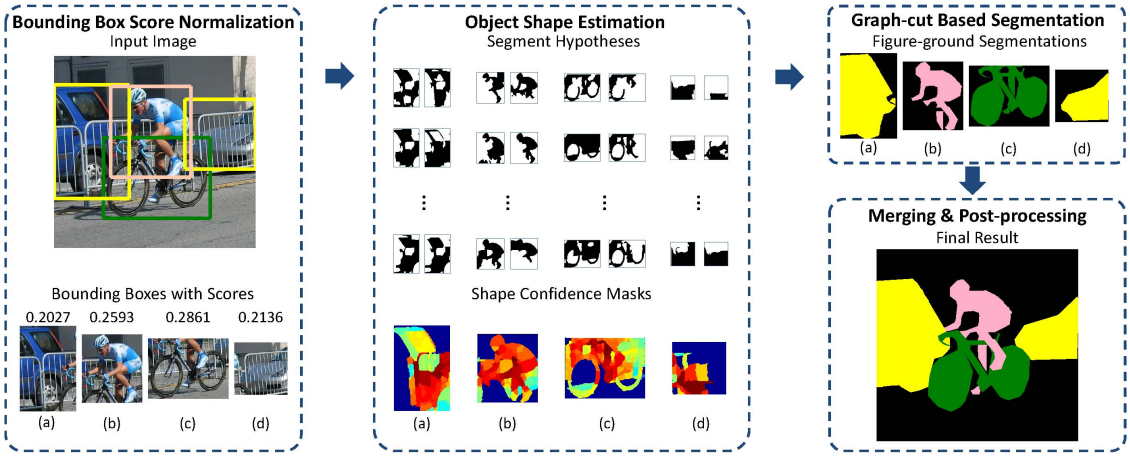


Figure 2. Overview of the proposed approach. First, the object bounding boxes with detection scores are extracted from the test image. Then, a voting based scheme is applied to estimate object shape guidance. By making use of the shape guidance, a graph-cut-based figure-ground segmentation provides a mask for each bounding box. Finally, these masks are merged and post-processed to obtain the final result.

in order to get the final segmentation results. The main challenge is to obtain object shape templates, especially for objects with relatively large intra-class appearance and pose variations.

Ladicky *et al*. [18] proposed a multilevel hierarchical *conditional random field* (CRF) model to incorporate information from different scales, which is combined with top-down detectors and global occurrence information [20]. Boix *et al*. [3] proposed so-called *harmony potential*, which integrates global category label information as well as object detectors in order to better fuse global and local information. Although CRF-based models have strong generalization capability to integrate different cues from different scales, the modelling and training of these kinds of methods are relatively difficult due to the large number of parameters.

In this paper, we propose an efficient, learning-free design for semantic segmentation when the object bounding boxes are available (see Fig. 1). Its key aspects and contributions (see Fig. 2) are summarized as below:

- In some situations, training data with annotated segments are not available, making learning based methods including the state-of-the-art CPMC-based frameworks [7] infeasible. However, the object bounding boxes can be obtained in a much easier way, either through user interaction or from object detector which also provides class label as additional information. Here, we propose an approach based on detected bounding boxes, where no additional segment annotation from the training set or user interaction is required.
- Shape information can substantially improve the segmentation [25]. However, to obtain shape information is sometimes quite challenging because of the large intra-class variability of the objects. Based on a set of segment hypotheses, we introduce a simple voting scheme to estimate the shape guidance. The derived shape guidance is used in the subsequent graph-cut-based formulation to provide the figure-ground segmentation.
- Comprehensive experiments on the most challenging object segmentation datasets [12, 22] demonstrate that the performance of the proposed method is competitive or even superior against to the state-of-the-art methods. We also conduct an analysis of the effect of the bounding box accuracy.

2. Related Work

Numerous semantic segmentation methods utilize the object bounding box as a prior. The bounding boxes are provided by either user interaction or object detectors. These methods tend to exploit the provided bounding box merely to exclude its exterior from segmentation. A probabilistic model is described in [27] that captures the shape, appearance and depth ordering of the detected objects on the image. This layered representation is applied to define a novel deformable shape support based on the response of a mixture of part-based detectors. In fact, the shape of a detected object is represented in terms of a layered, per-pixel segmentation. Dai *et al*. [11] proposed and evaluated several color models based on learned graph-cut segmentations to help re-localize objects in the initial bounding boxes predicted from deformable parts model (DPM) [13]. Xia *et al*. [26] formulated the problem in a sparse reconstruction framework pursuing a unique latent object mask. The objects are detected on the image, then for each detected bounding box, the objects from the same category along with their object masks are selected from the training set and transferred to a latent mask within the given bounding box. In [16] a principled Bayesian method, called OBJ

CUT, is proposed for detecting and segmenting objects of a particular class label within an image. This method [16] combines top-down and bottom-up cues by making use of object category specific Markov random fields (MRF) and provides a prior that is global across the image plane using so-called pictorial structures.

In [24], the traditional graph-cut approach is extended. The proposed method [24], called GrabCut, is an iterative optimization and the power of the iterative algorithm is used to simplify substantially the user interaction needed for a given quality of result. GrabCut combines hard segmentation by iterative graph-cut optimization with border matting to deal with blurred and mixed pixels on object boundaries. In [22] a method is introduced which further exploits the bounding box to impose a powerful topological prior. With this prior, a sufficiently tight result is obtained. The prior is expressed as hard constraints incorporated into the global energy minimization framework leading to an NP-hard integer program. The authors [22] provided a new graph-cut algorithm, called *pinpointing*, as rounding method for the intermediate solution.

In [9], an adaptive figure-ground classification algorithm is presented to automatically extract a foreground region using a user provided bounding box. The image is first oversegmented, then the background and foreground regions are gradually refined. Multiple hypotheses are generated from different distance measures and evaluation score functions. Finally, the best segmentation is automatically selected with a voting or weighted combination scheme.

3. Proposed Solution

In this section, we introduce the proposed solution in details. For a given test image, first the object bounding boxes with detection scores are predicted by object detectors. The detection scores are normalized and some bounding boxes with low scores are removed (see Section 3.1). A large pool of segment hypotheses are generated by purely applying CPMC method [8] (without using any learning process), in order to estimate the object shape guidance in a given bounding box. The shape guidance is then obtained by a simple but effective voting scheme (see Section 3.2). The derived object shape guidance is integrated into a graph-cut-based optimization for each bounding box (see Section 3.3). The obtained segmentation results corresponding to different bounding boxes are merged and further refined through some post-processing techniques including morphological operations, *e.g*. hole filling (see Section 3.4). The pipeline of the proposed approach is presented in Fig. 2.

3.1. Bounding Box Score Normalization

In order to obtain the bounding boxes, we apply the state-of-the-art object detectors provided by the authors of [10, 28]. For a given test image, class-specific object detectors provide a set of bounding boxes with class labels and detection scores. For interacting objects (*e.g*. bike and the human on Fig. 1), we need to compare the detection results over the overlapping areas. While comparing two objects taken from different classes, it is observed that the higher score does not necessarily mean the higher probability of being an object instance from the given class, since the score value scales are class-specific.

In order to transform the detection scores, we introduce some standardizing measures. The *precision* is the fraction of retrieved objects that are relevant and the *recall* is the fraction of relevant objects that are retrieved. The *F-measure* is defined as the harmonic mean of the precision and recall. By applying the different detection scores as threshold values over the objects in the validation set, one can estimate the *precision over score* (PoS) function for a given class. Since the values of the PoS function are only a function of the objects in the validation set, its piecewise linear approximation is pre-calculated over the interval $[0, 1]$.

By substituting the actual detection scores into PoS functions, one can transform and compare the scores provided by detectors from different classes. Nevertheless, for some score values, the corresponding precisions are too low making the PoS function unreliable. To overcome this problem, let r_c^* denote the recall value where the F-measure is maximal (*i.e*. the precision value is equal to the recall value) for a given class c. Those detection scores whose recall values are greater than r_c^* imply that the precision ($\leq r_c^*$) is not reliable enough. Hence we apply r_c^* as a threshold to restrict the domain of the PoS function relating to the class c to the interval $[r_c^*, 1]$, while leaving its value to be zero outside this domain.

In our experiments, the bounding boxes that have lower detection scores than a threshold value (τ) are removed. Note that we can use a common threshold value for all classes, since the detection scores are now comparable.

3.2. Object Shape Guidance Estimation

After obtaining object bounding boxes, a figure-ground segmentation is performed for each bounding box. As figure-ground segmentation methods [17, 15] can benefit significantly from the shape guidance, we introduce a simple yet effective idea to obtain the shape guidance. For this purpose, a set of object segments serving as various hypotheses for the object shape, is generated for the given test image. The object shape is then estimated based on a simple voting scheme.

The segment hypotheses are generated by solving a sequence of CPMC problems [8] without any prior knowledge about the properties of individual object classes. So, only the unsupervised part of [8] is applied here without any subsequent ranking or classification of the generated segments,

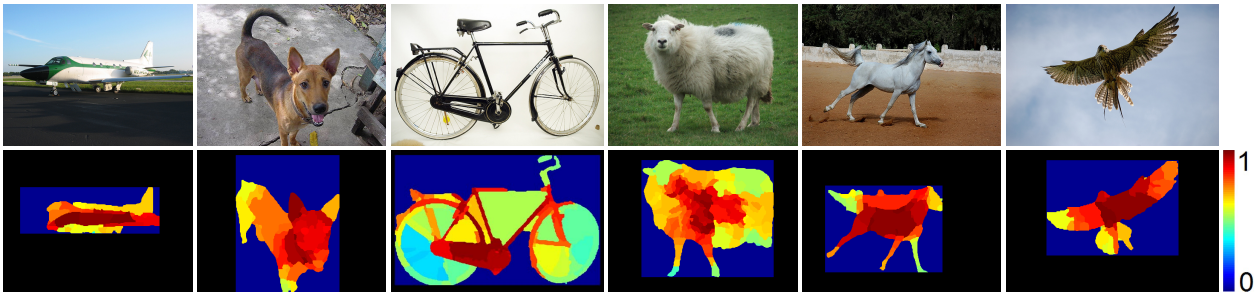


Figure 3. Some exemplar images (**top**) and the estimated object shape guidance with shape confidence (**bottom**). (Best viewed in color.)

hence no training annotation is needed. This method [8] provides visually coherent segments by varying the parameter of the foreground bias.

The information about the object localization is provided by the bounding box, and hence we can crop the segments. The small segments can be considered as noise whereas the very large ones usually contain a large portion of the background region. Therefore, we omit those segments smaller than $\gamma_1 = 20\%$ or larger than $\gamma_2 = 80\%$ of the bounding box area. Let $\mathcal{S}_1, \ldots, \mathcal{S}_k \subset \mathbb{R}^2$ denote the regions of the remaining cropped segments. Then the average map $\bar{M} : \mathbb{R}^2 \to \mathbb{R}$ is calculated for each pixel $\mathbf{p}$ as

$$\bar{M}(\mathbf{p}) = \frac{1}{k} \sum_{i=1}^{k} \mathbb{1}_i(\mathbf{p}) \ ,$$

where $\mathbb{1}_i : \mathbb{R}^2 \to \{0, 1\}$ is the characteristic function of $\mathcal{S}_i$ for all $i = 1, \ldots, k$. $\bar{M}$ can be considered as a score map, where each segment gives equal vote. Those regions sharing more overlapping segments and thus higher scores, have higher confidence to be the part of the object shape.

The generated segments partially cover the object, nevertheless, some segment among $\mathcal{S}_1, \ldots, \mathcal{S}_k$ may still be inaccurate, and thus decrease the reliability of the shape guidance. We select the best overlapping segment that aligns well to the object boundary. The main challenge lies in how to identify such segments. Let $\mathcal{M}_t = \{\mathbf{p} \in \mathbb{R}^2 \mid \bar{M}(\mathbf{p}) \geq t\}$, and then the "best" segment is estimated as the solution of the problem:

$$i^* = \underset{i \in \{1,\ldots,k\}}{\arg\max} \left\{ \max_{t \geq \mu \max(\bar{M})} \frac{|\mathcal{M}_t \cap \mathcal{S}_i|}{|\mathcal{M}_t \cup \mathcal{S}_i|} \right\} \ ,$$

where $\mu = 0.25$ ensures a minimal confidence in the selection. The final object shape guidance is achieved by restricting the domain of $\bar{M}(\mathbf{p})$ based on the "best" segment, more precisely $M(\mathbf{p}) = \bar{M}(\mathbf{p}) \mathbb{1}_{i^*}(\mathbf{p})$. This approach provides the shape guidance as well as the shape confidence score for each pixel. Some examples of the estimated shape guidance are shown in Fig. 3.

3.3. Graph-cut Based Segmentation

We follow popular graph-cut based segmentation algorithms [4, 8], where the image is modelled as a weighted graph $G = \{\mathcal{V}, \mathcal{E}\}$, that is, the set of nodes $\mathcal{V} = \{1, 2, \ldots, n\}$ consists of super-pixels, while the set of edges $\mathcal{E}$ contains the pairs of adjacent super-pixels. For each node $i \in \mathcal{V}$ a random variable x_i is assigned a value from a finite label set $\mathcal{L}$. An energy function is defined over all possible labellings $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathcal{L}^n$ [4]:

$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}} u_i(x_i) + \sum_{(i,j) \in \mathcal{E}} v_{ij}(x_i, x_j) \ . \qquad (1)$$

The first term u_i, called *data term*, measures the disagreement between the labellings $\mathbf{x}$ and the image. The second term v_{ij}, called *smoothness term*, measures the extent to which $\mathbf{x}$ is not piecewise smooth. The data term should be non-negative, and the smoothness term should be a metric. The segmentation is obtained by minimizing Eq. (1) via graph-cut [8].

The data term u_i involves a weighted combination of color distribution and shape information with the weight $\alpha \in [0, 1]$

$$u_i(x_i) = -\left\{ \alpha \log\left(A(x_i)\right) + (1 - \alpha) \log\left(S(x_i)\right) \right\} . \qquad (2)$$

It evaluates the likelihood of x_i taking on the label $l_i \in \mathcal{L} = \{0, 1\}$ according to *appearance term* A and *shape term* S, where 0 and 1 respectively represent the background and foreground.

Let $\mathcal{V}_f$ and $\mathcal{V}_b$ denote the initial seeds for foreground and background regions, respectively. $\mathcal{V}_f$ and $\mathcal{V}_b$ are estimated based on the ratio of their overlap with the estimated shape guidance $\mathcal{M} = \{\mathbf{p} \in \mathbb{R}^2 \mid M(\mathbf{p}) > 0\}$, obtained in Section 3.2. By introducing the notation $\mathcal{R}_i$ for the region of the i$^{\text{th}}$ super-pixel, we define $\mathcal{V}_f$ and $\mathcal{V}_b$ as

$$\begin{aligned} \mathcal{V}_f &= \{i \in \mathcal{V} : |\mathcal{R}_i \cap \mathcal{M}| > \delta_1 |\mathcal{R}_i|\} \ , \\ \mathcal{V}_b &= \{i \in \mathcal{V} : |\mathcal{R}_i \cap \mathcal{M}| < \delta_2 |\mathcal{R}_i|\} \ , \end{aligned}$$

where $\delta_1 = 0.2$ and $\delta_2 = 0.8$. The appearance term A is defined as

$$A(x_i) = \begin{cases} 1 & \text{if } x_i = 1 \text{ and } i \notin \mathcal{V}_b \\ 0 & \text{if } x_i = 1 \text{ and } i \in \mathcal{V}_b \\ 0 & \text{if } x_i = 0 \text{ and } i \in \mathcal{V}_f \\ p_b(x_i)/p_f(x_i) & \text{if } x_i = 0 \text{ and } i \notin \mathcal{V}_f \end{cases}$$

where $p_f(x_i)$ and $p_b(x_i)$ return the probabilities of x_i being foreground and background, respectively, for the i$^{\text{th}}$ super-pixel. The probabilities are computed based on colors for each pixel and the average value is calculated for a given super-pixel. In order to estimate the probability density functions over the seeds of $\mathcal{V}_f$ and $\mathcal{V}_b$, we apply Gaussian mixture model with five components.

M can be considered as a confidence map, since its value for each pixel is calculated based on the number of overlapping segments. The shape term $S(x_i = 1)$ for the i$^{\text{th}}$ super-pixel is simply calculated by the average value of M over the overlapping area with the given super-pixel. Then $S(x_i = 0) = 1 - S(x_i = 1)$ is readily obtained. Note that this shape term immediately incorporates the spatial difference between the super-pixels and the shape guidance $\mathcal{M}$.

The smoothness term penalizes different labels assigned to adjacent super-pixels:

$$v_{ij}(x_i, x_j) = [x_i \neq x_j] e^{-d(x_i, x_j)} \ ,$$

where $[x_i \neq x_j] = 1$, if $x_i \neq x_j$ and 0 otherwise. The function d computes the color and edge distance between neighbouring nodes for some $\beta \geq 0$:

$$d(x_i, x_j) = \max\left(\text{gPb}(x_i), \text{gPb}(x_j)\right) + \beta \left\| \mathbf{c}(x_i) - \mathbf{c}(x_j) \right\|^2, \quad (3)$$

where $\text{gPb}(x_i)$ returns the average of the values provided by edge detector globalPb [2] for each pixel belonging to the i$^{\text{th}}$ super-pixel and $\mathbf{c}(x_i)$ denotes the average RGB color vector over the given super-pixel.

3.4. Merging and Post-processing

After obtaining figure-ground segmentations for the bounding boxes, the results are projected back to the image and merged. In case of intersecting areas, the label with higher detection score is assigned to the given region. If the detection scores are equal to each other, then the larger region is retained.

In order to remove some artifacts, morphological hole filling is also applied. Finally, the super-pixels are further refined by using super-pixels extracted on a finer level (*i.e.* 300 super-pixels) that align better with the real object boundaries. On the finer scale, if a super-pixel has an overlap with the coarse segmentation result larger than $\gamma_3 = 80\%$, then its label will be set as the category of the coarse region.

4. Experimental Results

We conduct comprehensive experiments to demonstrate the performance of the proposed method and also present comparison with previous methods. Most of the experiments were run on the PASCAL VOC 2011, 2012 object segmentation datasets [12] consisting of 20 object classes and an additional background class, where the average image size is 473×382 pixels. This dataset [12] is among the most challenging datasets in the semantic segmentation field. The *Intersection over Union* (IoU) [12] measure is applied for quantitative evaluation.

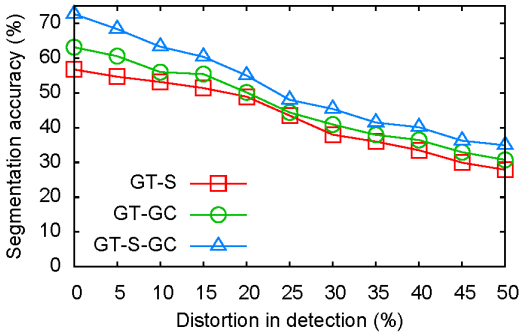


Figure 4. Effect of the distortion of bounding box.

In our experiments, we generated on average 547 segment hypotheses for each image by following [8]. For all images $n = 200$ super-pixels are obtained by [23]. The weights α and β (in Eq. (2) and Eq. (3)) are set to 0.5 according to cross validation experiments on VOC 2011 validation dataset. To solve graph-cut-based optimization we use the method in [4].

4.1. Proof of the Concept

In order to evaluate the impact of different parts of the proposed approach, a series of experiments has been conducted on the VOC 2011 validation dataset containing 1112 images. Note that ground truth bounding box information is also available for these images. We evaluated the quality of the segmentation results provided by the shape guidance $\mathcal{M}$ that is merged directly without running graph-cut optimization, referred as GT-S. GT-GC denotes the results obtained by the graph-cut formulation (Eq. (2)) when the shape guidance model is omitted ($\alpha = 1$). Finally, GT-S-GC denotes the case where $\alpha = 0.5$ is set in Eq. (2). We ran our proposed method with different settings, *i.e.* GT-S, GT-GC and GT-S-GC, for all images and obtained the average accuracy, calculated as the average of the IoU scores across all classes, 56.7%, 63.13% and 72.64%, respectively. The significant improvement from the GT-GC to GT-S-GC validates the effectiveness of the shape guidance in the proposed segmentation framework.

We have assumed that the bounding boxes provided by the object detectors are accurate enough, which is sometimes not the case. Here, we also analyze the effect of the bounding box accuracy. We evaluated the proposed method with different settings (GT-S, GT-GC and GT-S-GC) on various sets of bounding boxes with different accuracies. We should remark that the accuracy of object detectors is also evaluated by the IoU measure. Since the ground truth is given, we can generate new bounding boxes for each ob-

Table 1. Comparison of segmentation accuracy provided by previous methods on VOC 2011 test dataset [12].

Method	b/g	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	m/bike	person	plant	sheep	sofa	train	tv	avg
BONN-SVR [7]	**84.9**	**54.3**	**23.9**	39.5	35.3	42.6	65.4	53.5	**46.1**	**15.0**	**47.4**	**30.1**	33.9	48.8	**54.4**	46.4	28.8	**51.3**	26.2	44.9	37.2	43.3
BONN-FGT [14]	83.4	51.7	23.7	**46.0**	33.9	49.4	**66.2**	**56.2**	41.7	10.4	41.9	29.6	24.4	49.1	50.5	39.6	19.9	44.9	26.1	40.0	41.6	41.4
NUS-S	77.2	40.5	19.0	28.4	27.8	40.7	56.4	45.0	33.1	7.2	37.4	17.4	26.8	33.7	46.6	40.6	23.3	33.4	23.9	41.2	38.6	35.1
Brooks	79.4	36.6	18.6	9.2	11.0	29.8	59.0	50.3	25.5	11.8	29.0	24.8	16.0	29.1	47.9	41.9	16.1	34.0	11.6	43.3	31.7	31.3
Xia *et al*. [26]	82.3	48.2	23.2	38.7	**36.1**	49.0	62.4	40.6	39.6	13.1	38.4	21.6	37.8	49.7	48.4	**53.2**	25.5	36.0	31.5	46.8	48.8	41.5
Arbeláez *et al*. [1]	83.4	46.8	18.9	36.6	31.2	42.7	57.3	47.4	44.1	8.1	39.4	36.1	36.3	49.5	48.3	50.7	26.3	47.2	22.1	42.0	43.2	40.8
DET1-Proposed	83.4	51.2	23.4	40.6	32.4	**51.3**	63.5	52.8	44.9	14.2	45.8	20.2	**39.6**	**53.5**	51.7	45.4	**38.4**	44.5	**32.3**	**48.6**	**49.5**	**44.1**

Table 2. Comparison of segmentation accuracy provided by previous methods on VOC 2012 test dataset [12].

Method	b/g	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	m/bike	person	plant	sheep	sofa	train	tv	avg
O2P-CPMC-CSI	85.0	59.3	27.9	43.9	39.8	41.4	52.2	61.5	56.4	13.6	**44.5**	26.1	42.8	51.7	**57.9**	51.3	29.8	45.7	28.8	49.9	43.3	45.4
CMBR-O2P-CPMC-LIN	83.9	60.0	27.3	46.4	40.0	41.7	57.6	59.0	50.4	10.0	41.6	22.3	43.0	51.7	56.8	50.1	33.7	43.7	29.5	47.5	44.7	44.8
O2P-CPMC-FGT-SEGM	**85.1**	**65.4**	29.3	**51.3**	33.4	44.2	59.8	60.3	52.5	13.6	53.6	**32.6**	40.3	57.6	57.3	49.0	33.5	53.5	29.2	47.6	37.6	47.0
NUS-DET-SPR-GC-SP	82.8	52.9	31.0	39.8	44.5	58.9	60.8	52.5	49.0	**22.6**	38.1	27.5	47.4	52.4	46.8	51.9	35.7	**55.3**	**40.8**	**54.2**	47.8	47.3
UVA-OPT-NBNN-CRF	63.2	10.5	2.3	3.0	3.0	1.0	30.2	14.9	15.0	0.2	6.1	2.3	5.1	12.1	15.3	23.4	0.5	8.9	3.5	10.7	5.3	11.3
DET2-Proposed	82.9	49.1	**30.5**	44.6	**36.6**	59.5	**65.7**	53.0	51.9	21.8	41.5	25.0	44.9	54.7	49.4	49.6	33.2	49.6	37.5	53.1	48.7	46.8
DET3-Proposed	82.5	52.1	29.5	50.6	35.6	**59.8**	64.4	**55.5**	**54.7**	22.0	38.7	24.3	**48.3**	**55.6**	52.9	**52.2**	**38.2**	49.1	35.5	53.7	**53.5**	**48.0**

Figure 5. The most common cases of mis-detection of the objects due to rare pose, cluttered background and occlusion.

ject in the validation dataset by modifying the corner points of the bounding boxes. Thus, we randomly modified the ground truth bounding boxes based on uniform distribution to achieve $5\%, 10\%, \ldots, 50\%$ distortions in the accuracy. Fig. 4 shows the performance of the different settings of the proposed method for given distortions in detection. As can be seen on Fig. 4, more accurate bounding boxes lead to better performance in segmentation, since it provides not only more accurate localization, but also more accurate cropped segments to estimate the shape guidance. Furthermore, the shape guidance term provides important top-down guidance prior that improves the final results.

4.2. Comparison with the State-of-the-arts

Here, we present a comprehensive comparison with the state-of-the-arts. The PoS functions for different object classes were estimated on the detection validation dataset, which is also available in [12]. The threshold value for the bounding boxes τ is set to 0.2 based on cross-validation.

VOC 2011 test dataset Table 1 shows the detailed comparison of the proposed method with previous approaches on the VOC 2011 segmentation challenge. Among the competing methods, BONN-SVR [7] and BONN-FGT [14] also utilize detection annotations in the training stage. The methods NUS-S and Brooks apply CRF-based framework to integrate information cues from different levels. Xia *et al*. [26] and Arbeláez *et al*. [1] are two state-of-the-art detection based methods. The results of the proposed method are obtained by applying the state-of-the-art object detector [10, 28], referred as DET1-Proposed.

It can be seen from Table 1 that our proposed method achieves superior results as compared to both other detection based methods and the VOC 2011 winner BONN-SVR [7]. Among the 21 classes including the background, DET1-Proposed achieves the best performance in 7 classes with an average accuracy of 44.1%, which is 0.8% higher than that of the VOC 2011's winner.

To the best of our knowledge, this is the best result reported on this dataset, when all the training data are strictly from the VOC 2011 dataset. Note that in this work, we do not mean to claim that our method is always superior over CPMC-based method [7]. It is predictable that, the CPMC-based method [7] could achieve better results with more annotated data or more accurate detection information. For instance, [6] reported a better performance of 47.6% by using extra annotation data besides the VOC 2011 training set (more than 13000 images with ground truth semantic edge annotation) to train the model. However, the proposed unsupervised framework is competitive even without annotated segments from either the training set or external dataset.

VOC 2012 test dataset Table 2 shows the detailed comparison of the proposed method to top-performing algorithms on the latest VOC 2012 segmentation challenge. Easy to observe that almost all methods are combination of previous methods. The first three competing methods are based on CPMC [8]. O2P-CPMC-CSI utilizes a novel probabilistic inference procedure, called composite statistical inference (CSI), in which the predictions of overlapping figure-ground hypotheses are used. CMBR-O2P-CPMC-LIN applies a simple linear SVR with second order pooling [6]. O2P-CPMC-FGT-SEGM is based on the original BONN-SVR [14, 6] approach. UVA-OPT-NBNN-CRF applies a CRF-based framework with naive Bayes nearest neighbour (NBNN) features. NUS-DET-SPR-GC-SP is the VOC 2012 winner, which is also a detection based method

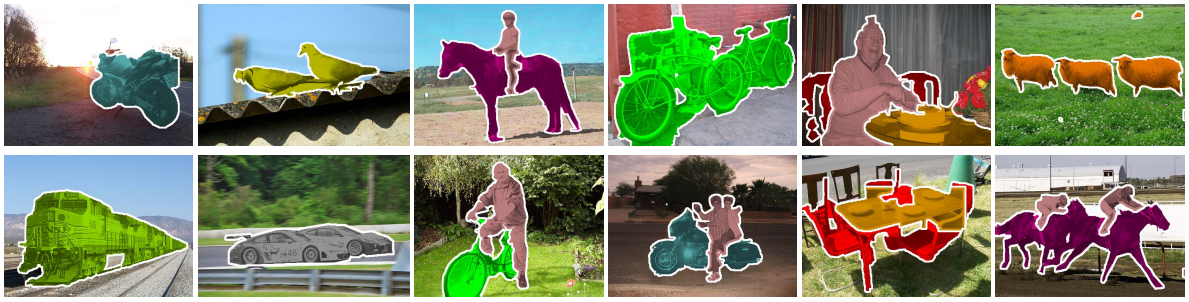


Figure 6. Some exemplar results on VOC 2012 test dataset [12] provided by our proposed method (DET3-Proposed). The results are overlaid on the images with white boundaries and different colors correspond to different categories. (Best viewed in color.)

based on [26] followed by an MRF refinement process.

For some images, however, the current state-of-the-art object detector in [10, 28] (referred as DET1) cannot provide bounding boxes with higher score than τ leading to mis-detection of the objects. This is often due to rare pose, cluttered background and occlusion (see Fig. 5). As demonstrated in Section 4.1, increasing detection accuracy will improve the segmentation performance. Therefore to further validate our claim in practical cases, we designed a boosted "object detector" (referred as DET2). DET2 predicts some bounding boxes based on segmentation results obtained from [3] only for the images without bounding box prediction from DET1, otherwise, the bounding boxes from DET1 are considered. DET3 directly obtains bounding boxes from segmentation results of NUS-DET-SPR-GC-SP, which is our submitted methods to VOC 2012 challenge.

The results in Table 2 show that DET2-Proposed performs the best in 3 out of the 21 categories while DET3-Proposed performs the best in 8 out of the 21 the categories, which is the highest among all the competing methods. Furthermore, DET3-Proposed achieves the best average performance of 48%. Note that only the estimated bounding boxes are used in our solution, which contain much less information than the segmentation results, hence the improvement of 0.7% from NUS-DET-SPR-GC-SP (47.3%) is reasonable. Although DET2 and DET3 implicitly use ground truth segments which seems to contradict with our claim that no annotated segments are needed, we aim to further validate that better detection leads to better segmentation (see Section 4.1) in practical cases. DET2 and DET3 just demonstrate the potential improvement when more accurate detector is available.

Some qualitative results are shown in Fig. 6 containing images with single object as well as images with multiple interacting objects with rigid transportation tools, articulated animals and indoor objects. Based on these results, it is fair to say that the proposed method can well handle background clutters, objects with low contrast with the background and multiple objects, as far as the detection is accurate enough. However, there are some failure cases mainly due to mis-detection and inaccurate bounding box prediction or wrong class labelling (see Fig. 7).

Figure 7. Some failure cases obtained by the proposed method (DET3-Proposed). The results are overlaid on the images with white boundaries and different colors correspond to different categories. (Best viewed in color.) The first image is due to mis-detection of the small horse. The second one is due to wrong bounding box prediction, since the cloth is labelled as person and the parrot (bird) is mis-detected. The third one is due to inaccurate bounding box prediction (*i.e.* wrong label for the bottle) resulted in inaccurate estimation in the graph-cut formulation.

GrabCut-50 dataset We also compare the proposed method to the related segmentation frameworks guided by bounding box prior [24, 22, 9]. For this sake, these experiments were run on the GrabCut-50 [22] dataset consisting of 50 images with ground truth bounding boxes. The error-rate (denoted by ϵ) is computed as the percentage of mislabeled pixels inside the bounding box.

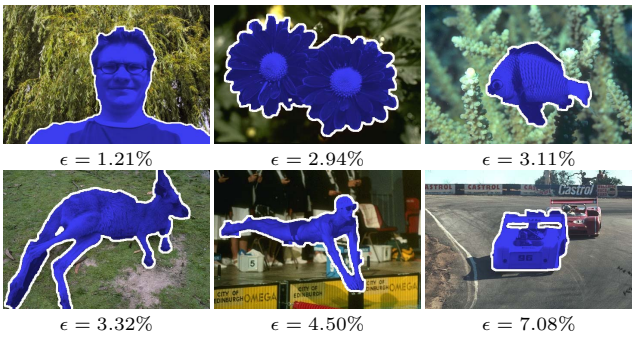
In these experiments, we generated the segment hypotheses for the whole image instead of the object bounding boxes. 400 and 800 super-pixels are extracted for graph-cut optimization and super-pixel refinement, respectively. In post-processing, the threshold γ_3 is set to 0.4 due to the much smaller size of the finer-scale super-pixels compared to the settings in the PASCAL VOC experiments. Finally, we applied morphological filtering (*i.e.* morphological opening and closing), instead of hole filling.

The results are shown in Table 3. Compared to the state-of-the-art methods CrabCut [24], GrabCut-Pinpoint [22] and F-G Classification [9], it is evident that the proposed method is superior in its better performance. GrabCut-Pinpoint uses an iterative solution and relies on the assumption that the bounding box is tight, which is not always true.

Some qualitative results are shown in Fig. 8. Note that this dataset [22] is easier than the VOC dataset [12] and contains only 50 images with single object in each image. The proposed method provides the error $\epsilon = 7.08\%$ in the worst

Table 3. Comparison with bounding box prior based algorithms on GrabCut-50 dataset.

Method	Error-rate ϵ
GrabCut [24]	8.1%
GrabCut-Pinpoint [22]	3.7%
F-G Classification [9]	5.4%
Proposed method	**3.3%**



$\epsilon = 1.21\%$ $\epsilon = 2.94\%$ $\epsilon = 3.11\%$

$\epsilon = 3.32\%$ $\epsilon = 4.50\%$ $\epsilon = 7.08\%$

Figure 8. Some segmentation results, overlaid on the images with blue color and white boundary, on the GrabCut-50 dataset [22] obtained by the proposed method.

case (see the last image on Fig. 8), which means that the performance is almost saturated in this dataset [22]. This also concludes that better bounding box prior significantly improves the final segmentation results.

Furthermore, we ever ran the DET1+GrabCut method as a baseline on the VOC 2011 dataset, and obtained the accuracy of 37.2%, which is much lower than our 44.1%. Therefore the superiority of the proposed framework over Grab-Cut [24] is further validated.

5. Conclusions

In this paper, we proposed a detection based learning free approach for semantic segmentation without the requirement of any annotated segments from the training set. Furthermore, a simple voting scheme based on a generated pool of segment hypotheses, is proposed to obtain the shape guidance. Finally graph-cut-based formulation is used to perform semantic segmentation. Extensive results on the challenging VOC 2011 and VOC 2012 segmentation datasets as well as the GrabCut-50 dataset demonstrate the effectiveness of the proposed framework.

Some general observations from the results are that the proposed method performs nearly perfectly in those cases with single object, while for images with multiple objects or interacting objects, the performance depends on the accuracy of the bounding box. Therefore, one of the main limitations of this approach is that the object detector inherently affects the segmentation performance. However, when no training data is available but the detection is given, this approach could act as a valid alternative approach for semantic segmentation.

With better object detectors, such as one that could well handle partial objects and occlusions, huge improvement could be expected for object segmentation performance. In addition, better ways to obtain the shape guidance and handle multiple interacting segments are also worth exploring to further refine the existing detection-based segmentation methods.

Acknowledgment This work is partially supported by the Singapore PSF grant 1321202075 and Singapore Ministry of Education under research grant MOE2010-T2-1-087.

References

[1] P. Arbeláez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, and J. Malik. Semantic segmentation using regions and parts. In *CVPR*, 2012.

[2] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *TPAMI*, 33(5):898–916, May 2011.

[3] X. Boix, J. Gonfaus, J. Weijer, A. Bagdanov, J. Gual, and J. Gonzàlez. Harmony potentials - fusing global and local scale for semantic image segmentation. *IJCV*, 96(1):83–102, Jan. 2012.

[4] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. In *ICCV*, 1999.

[5] T. Brox, L. Bourdev, S. Maji, and J. Malik. Object segmentation by alignment of poselet activations to image contours. In *CVPR*, 2011.

[6] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*, 2012.

[7] J. Carreira, F. Li, and C. Sminchisescu. Object recognition by sequential figure-ground ranking. *IJCV*, 98(7):243–262, July 2012.

[8] J. Carreira and C. Sminchisescu. CPMC: Automatic object segmentation using constrained parametric min-cuts. *TPAMI*, 34(7):1312–1328, July 2012.

[9] Y. Chen, A. Chan, and G. Wang. Adaptive figure-ground classification. In *CVPR*, 2012.

[10] Y. Chen, L. Zhu, and A. Yuille. Active mask hierarchies for object detection. In *ECCV*, 2010.

[11] Q. Dai and D. Hoiem. Learning to localize detected objects. In *CVPR*, 2012.

[12] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge (VOC). http://pascallin.ecs.soton.ac.uk/challenges/voc/.

[13] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IJCV*, 32(9):1627–1645, 2010.

[14] A. Ion, J. Carreira, and C. Sminchisescu. Probabilistic joint image segmentation and labeling. In *NIPS*, 2011.

[15] J. Kim and K. Grauman. Shape sharing for object segmentation. In *ECCV*, 2012.

[16] P. Kumar, P. Torr, and A. Zisserman. Obj Cut. In *CVPR*, 2005.

[17] D. Küttel and V. Ferrari. Figure-ground segmentation by transferring window masks. In *CVPR*, 2012.

[18] L. Ladický, C. Russell, P. Kohli, and P. Torr. Associative hierarchical CRFs for object class image segmentation. In *ICCV*, 2009.

[19] L. Ladický, C. Russell, P. Kohli, and P. Torr. Graph cut based inference with co-occurrence statistics. In *ECCV*, 2010.

[20] L. Ladický, P. Sturgess, K. Alahari, C. Russell, and P. Torr. What, where and how many? Combining object detectors and CRFs. In *ECCV*, 2010.

[21] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[22] V. Lempitsky, P. Kohli, C. Rother, and T. Sharp. Image segmentation with a bounding box prior. In *ICCV*, 2009.

[23] G. Mori, X. Ren, A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. In *CVPR*, 2004.

[24] C. Rother, V. Kolmogorov, and A. Blake. "GrabCut": interactive foreground extraction using iterated graph cuts. *TOG*, 23(3):309–314, Mar. 2004.

[25] N. Vu and B. Manjunath. Shape prior segmentation of multiple objects with graph cuts. In *CVPR*, 2008.

[26] W. Xia, Z. Song, J. Feng, L.-F. Cheong, and S. Yan. Segmentation over detection by coupled global and local sparse representations. In *ECCV*, 2012.

[27] Y. Yang, S. Hallman, D. Ramanan, and C. Fowlkes. Layered object detection for multi-class segmentation. In *CVPR*, 2010.

[28] L. Zhu, Y. Chen, A. Yuille, and W. Freeman. Latent hierarchical structural learning for object detection. In *CVPR*, 2010.