# Towards Unified Object Detection and Semantic Segmentation

Jian Dong[1], Qiang Chen[1], Shuicheng Yan[1] and Alan Yuille[2]

[1] Department of Electrical and Computer Engineering, NUS, Singapore
[2] Department of Statistics, UCLA, Los Angeles, CA, USA

**Abstract.** Object detection and semantic segmentation are two strongly correlated tasks, yet typically solved separately or sequentially with substantially different techniques. Motivated by the complementary effect observed from the typical failure cases of the two tasks, we propose a unified framework for joint object detection and semantic segmentation. By enforcing the consistency between final detection and segmentation results, our unified framework can effectively leverage the advantages of leading techniques for these two tasks. Furthermore, both local and global context information are integrated into the framework to better distinguish the ambiguous samples. By jointly optimizing the model parameters for all the components, the relative importance of different component is automatically learned for each category to guarantee the overall performance. Extensive experiments on the PASCAL VOC 2010 and 2012 datasets demonstrate encouraging performance of the proposed unified framework for both object detection and semantic segmentation tasks.

**Keywords:** Object Detection, Semantic Segmentation, Unified Approach

## 1 Introduction

Object detection and semantic segmentation are two core tasks of visual recognition [13, 19, 36, 3, 41, 6, 35]. Object detection is often formulated as predicting a bounding box enclosing the object of interest [19] while semantic segmentation usually aims to assign a category label to each pixel from a pre-defined set [6]. Though strongly correlated, these two tasks have typically been approached as separate problems and handled using substantially different techniques.

Template based detection using sliding window scanning (*e.g.* HoG [13] and DPM [19]) has long been the dominant approach for object detection. Though good at finding the rough object positions, this approach usually fails to accurately localize the whole object via a tight bounding box. In fact, it has been found that the largest source of detection error is inaccurate bounding box localization ($0.1 \leq \text{overlap} < 0.5$) [12, 25]. This may arise from the limited representation ability of template-based detectors for non-rigid objects. For example, the deformable part-based model (DPM) [19] detector works much better for localizing rigid cat heads than for more amorphous cat bodies [32]. As shown in

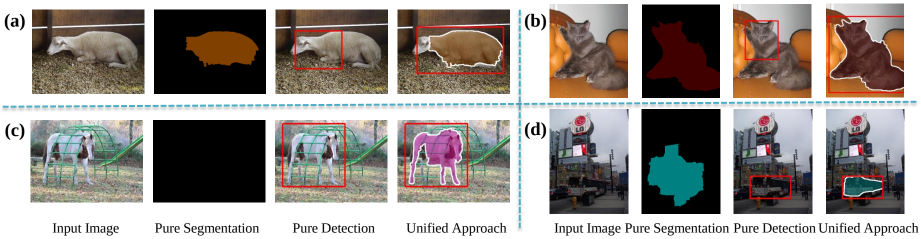


**Fig. 1.** The inconsistency of failure cases for object detection and semantic segmentation. The images in the top row show the scenario where detection is imperfect due to pose variance while the semantic segmentation works fine. The images in the bottom row show the scenario where semantic segmentation is not accurate while detectors can easily locate the objects. Thus, the two tasks are able to benefit each other, and more satisfactory results can be achieved for both tasks using our unified framework.

Figure 1 (a) and (b), the DPM detector often locates the head region only, which leads to the localization error. On the other hand, owing to their homogeneous appearances, the whole objects (cat and sheep) can be easily segmented out by the leading semantic segmentation techniques [6]. If poor localizations can be corrected with the help of semantic segmentation techniques [6], the overall detection performance would be improved considerably from additional true positives and fewer false positives.

Hypotheses based semantic segmentation has achieved great success during the past few years, which works by directly generating a pool of segment hypotheses for further ranking [2, 6]. However, due to the lack of global shape models, these approaches may fail to recognize the hypotheses of objects with heterogeneous appearances in the cluttered background, especially when all the generated hypotheses have some artifacts. As shown in Figure 1 (c) and (d), the leading hypotheses based semantic segmentation approach [6] either fails to segment out the object of interest or selects a much larger segment hypothesis. In contrast, if the target object has strong shape cues, the template-based detector [19] can easily locate the object and thus provide valuable information for semantic segmentation. Recently, a line of works, called detection-based segmentation, explored directly utilizing the detection results as top-down guidance and then performing segmentation within the given bounding boxes [4, 37]. However, such approaches usually have to make a hard decision about detection results at the early stage. Hence the error for detection, especially the localization error, will propagate to the segmentation results and could not be rectified. Intuitively it is beneficial to postpone making a hard decision till the last step of the pipeline [38].

Based on the above observations, we argue that object detection and semantic segmentation should be addressed jointly. Object detections should be consistent with some underlying segments to integrate local cues for better localization as shown in Figure 1 (a) and (b). Similarly, hypotheses based semantic segmentation should benefit from template-based object detectors to select better segment hypotheses as shown in Figure 1 (c) and (d). To this end, we propose a principled framework to unify current leading object detection and semantic segmentation techniques. By enforcing the consistency, our unified approach can benefit from the advantages of both techniques. In addition, some ambiguous object hypotheses may be difficult to classify from the information within the window/segment alone, but contextual information, such as local context around each object hypothesis and global image-level context, can help [29, 34, 11]. Hence, we further integrate contextual modeling into our framework. The major contributions of this work can be summarized as follows:

- We propose a principled framework for joint object detection and semantic segmentation. By enforcing the consistency between detection and segmentation results, our unified framework can effectively leverage the advantages of both techniques. Furthermore, both local and global context information are integrated into our unified framework to distinguish the ambiguous examples.
- With our unified framework, all information is accumulated at the final stage of the pipeline for decision making. Hence, it is avoided to make any hard decision at the early stage. The relative importance of different components is automatically learned for each category to guarantee the overall performance.
- Extensive experiments are conducted for both object detection and semantic segmentation tasks on the PASCAL VOC [17] datasets. The state-of-the-art performance of the proposed framework verifies its effectiveness, showing that performing object detection and semantic segmentation jointly is beneficial for both tasks.

## 2 Related Work

Recently, by noticing the limitation and complementarity of techniques for both tasks, some researchers have begun to investigate their correlations [26, 2, 5, 39]. The early work [26] simply employs the masks from detectors to initialize graph-cuts based segmentations. In [28, 37], more sophisticated models are proposed to refine the region within ground-truth bounding boxes. Rather than focusing on entire objects, Brox *et al.* employed Poselet detectors to predict masks for object parts [5]. Arbeláez *et al.* aggregated top-down information from detectors as activation features for bottom-up segments [2]. Conversely, segmentation techniques have also been explored to assist object detection in different ways. Dai *et al.* utilized segments extracted for each object detection hypothesis for better localization [12]. Fidler *et al.* [20] proposed to improve object detection based on semantic segmentation results [6]. The segments and detection windows are associated with several manually designed geometry features. Unfortunately, nearly all the above approaches utilize a sequential manner to fuse detection and segmentation techniques. Hence, the overall performance heavily relies on

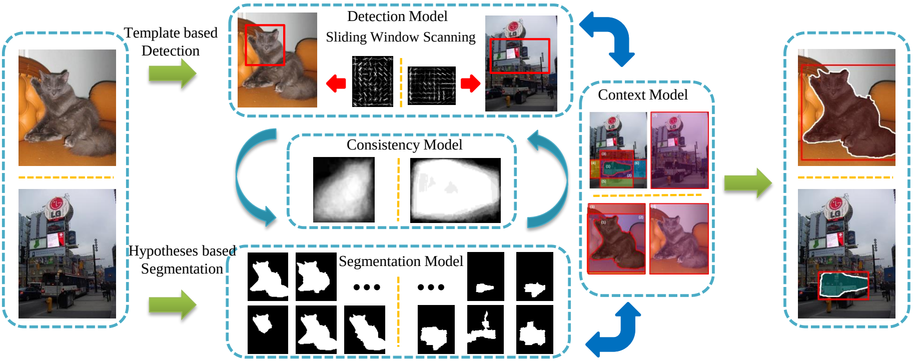


**Fig. 2.** Overview of the proposed unified object detection and semantic segmentation framework. Give a testing image, our UDS framework performs template based detection using sliding window scanning and hypotheses based semantic segmentation jointly. The agreement of the predictions from these two approaches is ensured by the consistency model. Both local context around the object hypothesis and global image context are also seamlessly integrated into our framework. The final output is the bounding box position and the index of the selected segment hypothesis.

the correctness of the initial results as the errors in the early stage are difficult to rectify.

Probably the most similar approach to ours is [27], which also aims to perform joint object detection and semantic segmentation. Our framework is different in the sense that we avoid making any hard decision at the early stage. All the information is aggregated at the final stage of the pipeline for decision making. On the contrary, [27] has to make initial decision about detection results. Hence, the initial detection errors, such as localization error, are difficult to rectify. Furthermore, unlike the CRF based model used in [27], we employ a hypotheses based approach for semantic segmentation. Hence, it is easier to ensure the shape consistency of top-down and bottom-up information in our framework.

## 3 Unified Object Detection and Semantic Segmentation

In this section, we introduce the details of the proposed unified object detection and semantic segmentation (UDS) framework. We start with an overview of the system and then detail each key component.

Figure 2 illustrates the pipeline of the proposed UDS framework. For the segmentation component, we employ the hypotheses based approach. Thus, with a pool of generated segment hypotheses, the segmentation problem is converted into choosing the appropriate hypothesis. Given a testing image, we perform template based detection using sliding window scanning and hypotheses based semantic segmentation jointly. Successful detection and segmentation require the agreement of both detection and segmentation predictions, which is achieved by utilizing a consistency model. In addition, as context plays an important role in distinguishing ambiguous object hypotheses, we further design a context model to aggregate both local (around the target object) and global (image-level) context information. For different object categories, each of these four components may have a different level of importance, which is automatically decided during the learning process. The final output of our system is the bounding box position ($p_0$) and the selected segment index ($id$) for the target object.

Formally, the joint detection and segmentation is achieved via the maximization of the following score function:

$$\begin{aligned} S(I, z, id) = \lambda^{Dt} S^{Dt}(z|w^{Dt}, I) + \lambda^{Sg} S^{Sg}(id|w^{Sg}, I) \\ +\lambda^{Ct} S^{Ct}(z, id|w^{Ct}, I) + S^{Cs}(z, id|w^{Cs}), \end{aligned} \tag{1}$$

where $w^{Dt}$, $w^{Sg}$, $w^{Ct}$ and $w^{Cs}$ are the parameters for detection, segmentation, context and consistency component, respectively. $\lambda^{Dt}$, $\lambda^{Sg}$, $\lambda^{Ct}$ are scalar weights for the corresponding components. $z$ captures the information for the template based detector and $id$ denotes the index of the selected segment. The details of each component are introduced in the following subsections. Based on the proposed unified approach, we avoid making any hard decision at the early stage. The final decision is delayed to the last step of the pipeline with all the integrated information, which implicitly relies on the learning mechanism to assess the relative importance of different components for each object category to guarantee the overall performance.

Finally, we want to emphasize that the proposed UDS framework provides a principled way to unify detection and segmentation techniques. We can directly employ the existing techniques or design new approaches for each component. Hence, it is easy to tailor UDS for specific applications, such as simultaneous person detection and segmentation. In this work, we will focus on utilizing the UDS framework for general object detection and semantic segmentation to verify its effectiveness.

### 3.1 Template based Detection Component

For the detection component, we aim to utilize the template based approach [19, 14], as it is good at capturing the shape cue and thus complementary to the appearance based segmentation techniques [6, 38]. In addition, through the mixture model strategy [19], these approaches can easily encode sub-category level top-down information (subcategory specific soft shape mask in this work). In this paper, we utilize the state-of-the-art deformable part-based model (DPM) [19]. Following [19], we define $z = \{c, p\}$, where $p = \{p_i\}_{i=0,\cdots,m}$. Here, $c$ denotes the mixture component index. $p_0$ encodes the location and scale of the root bounding box in an image pyramid and $\{p_i\}_{i=1,\cdots,m}$ encodes the $m$ part bounding boxes at the double resolution of the root. By concatenating the parameters for all mixtures as in [19], the score of a configuration can be written as

$$S^{Dt}(p, c|w^{Dt}, I) = \sum_{i=0}^{m} w_i^{Dt} \cdot \phi^{Dt}(I, p_i, c) + \sum_{i=1}^{m} w_{i,def}^{Dt} \cdot \phi^{Dt}(p_0, p_i, c), \qquad (2)$$

where $\phi^{Dt}(I, p_i, c)$ and $\phi^{Dt}(p_0, p_i, c)$ are the HoG pyramid features and spring deformation features, respectively, as in [19]. As Eqn. (2) is linear in model parameters, it can be written compactly as:

$$S^{Dt}(p, c|w^{Dt}, I) = w^{Dt} \cdot \phi^{Dt}(I, p, c). \qquad (3)$$

### 3.2 Hypotheses based Segmentation Component

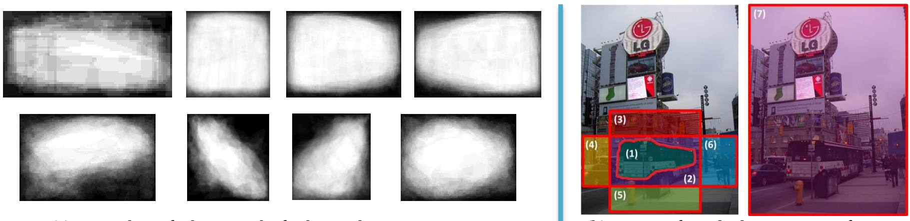
Hypotheses based semantic segmentation has achieved great success during the past few years [7, 6, 38]. This line of approaches mainly consist of two stages. The first stage generates a pool of segment hypotheses. The second stage ranks the generated hypotheses based on category-dependent information. The top ranked segments are returned as the final solution. Many efforts have been devoted to hypotheses generation through either a pure bottom-up approach [7, 36, 2] or a CRF based approach [38]. For the second stage, most approaches [7, 36, 38] simply employ the appearance based classification/regression for ranking. However, due to the limited discriminative ability of the appearance based ranking function, there exists a large gap between upper-bound accuracy of generated hypotheses (larger than 80%) and predicted accuracy of selected hypotheses (less than 50%) [7, 38]. As shown in Figure 1, due to the lack of global shape models, semantic segmentation relying on pure appearance based ranking may fail to find the appropriate hypotheses.

Based on the above observation, it may be expected that considerable improvement over the current segmentation performance can be achieved by means of simply selecting better hypothesis without generating more hypotheses. Hence, in this work we use standard methods for hypotheses generation and focus on selecting better segment hypotheses. To allow direct comparison, we utilize the publicly available code of the second order pooling ($O_2P$) approach [6] for hypotheses generation. For the feature representation $\phi^{Sg}(I, id)$ of the selected hypothesis $id$, a naive strategy is directly employing the second order pooling features as in [6]. However, training a latent model with high dimension features may be intractable. Hence, rather than keeping $\phi^{Sg}(I, id)$ as a high dimensional vector of raw second order pooling features, we represent $\phi^{Sg}(I, id)$ as the scores of pre-trained support vector regressors (SVR) [6]. Then, the score function of the segmentation component can be written as:

$$S^{Sg}(id|w^{Sg}, I) = w^{Sg} \cdot \phi^{Sg}(I, id). \qquad (4)$$

### 3.3 Consistency Component

The consistency component mainly aims to enforce the consistency between detection and segmentation prediction and thus leverage the advantages of both



(a) Exemplar soft shape masks for bus and cat categories

(b) Regions for calculating context features

**Fig. 3.** (a) Examples of subcategory-specific soft shape masks for buses (top row) and cats (bottom row). (b) Illustration of regions defined for computing the context features. Based on the selected segment hypothesis and bounding box, we adaptively divide the image into 7 regions as described in Section 3.4.

approaches. Soft shape mask has demonstrated to be effective for many detection guided techniques [39, 2, 8]. Hence, in this work, we measure the consistency between results of detection and segmentation approaches by calculating the correlations between their masks as shown below:

$$S^{Cs}(z, id|w^{Cs}) = \sum_{i=0}^{m} w_i^{Cs} \cdot m(p_i, id, c) = w^{Cs} \cdot \phi^{Cs}(p, id, c), \tag{5}$$

where $m(p_i, id, c)$ is the binary map $\{1, -1\}$ clipped from the segmentation hypothesis $id$ by the localized bounding box $p_i$. Here, $c$ in $m(p_i, id, c)$ is only used for padding 0 to make the equation with mixture models more compact, which is a common trick for the DPM approach [19].

Intuitively, the learned soft mask $w^{Cs}$ from top-down detection techniques can be seen as a shape guidance for bottom-up segmentation techniques. Enforcing the correlation between masks from both approaches will guarantee the consistency of top-down and bottom-up information. In addition, the mixture model strategy is critical to cope with variance in the poses as well as the view points. To ensure obtaining a reliable shape mask for each mixture component, we employ a shape guided mixture initialization as introduced in Section 4.2. Some examples of such soft shape masks are visualized in Figure 3 (a).

### 3.4 Context Component

Both the local context around the target object [29] and the global image context [34, 2, 11] have shown to be effective for visual recognition. The local context directly models the interaction of the target object and the surrounding environment. For example, a horse is often occluded by a person riding on it. In contrast, the global context mainly captures the image level information and co-existence/exclusion relation between objects.

In order to leverage such informative context cues, we further enhance the framework with an adaptive context model. Specifically, given a bounding box $p_0$

and a segment $id$, we divide the image into 7 regions (segment region, surrounding region within $p_0$, 4 context boxes and the whole image) as shown in Figure 3 (b). The area of the context box is half of that of the bounding box $p_0$. Hence, the spatial extent of the local context will vary adaptively based on $p_0$. If a context box crosses the boundary of the image, we consider only the area within the image. Fisher Vector (FV) [21, 9] is employed as region feature representation, as it has demonstrated the state-of-the-art performance for both object classification and detection [10, 11]. Furthermore, the average pooling strategy for FV enables effective calculation by utilizing the integral graph. Thus, the raw context representation is the concatenation of FVs on the 7 regions mentioned above.

Similar to the segmentation component, the dimension of the raw context features is too high. Hence, we first train a separate classifier for each object category and then use the predicted scores as the final context features. Then, the context component can be written as:

$$S^{Ct}(z, id|w^{Ct}, I) = S^{Ct}(p_0, id|w^{Ct}, I) = w^{Ct} \cdot \phi^{Ct}(I, id, p_0), \tag{6}$$

where $\phi^{Ctx}(I, id, p_0)$ is the concatenation of predicted scores for all classifiers. In fact, our context model can be seen as a variant of the appearance based detection approach to some extent. We still call it "context model" as it can provide valuable and complementary context information to the other three components.

## 4 Inference and Learning

This section introduces inference and learning of the proposed UDS framework. We begin with the general inference and learning procedure and then describe the implementation details in practice.

### 4.1 Inference

Similar to DPM [19], we employ the sliding windows strategy for inference. For a fixed root bounding box position $p_0$ and mixture index $c$, inference in our model can be done by solving the following optimization problem:

$$\begin{aligned} S(p_0, c) = &\max_{p_1, \cdots, p_m, id} S(p, id, c) = \max_{id} [\lambda^{Dt} w_0^{Dt} \cdot \phi^{Dt}(I, p_0, c) \\ &+ \lambda^{Sg} w^{Sg} \cdot \phi^{Sg}(I, id) + \lambda^{Ct} w^{Ct} \cdot \phi^{Ct}(I, id, p_0) + w_0^{Cs} \cdot m(p_0, id, c) \\ &+ \max_{p_1, \cdots, p_m} \sum_{i=1}^{m} (\lambda^{Dt} w_i^{Dt} \cdot \phi^{Dt}(I, p_i, c) + \lambda^{Dt} w_{i,def}^{Dt} \cdot \phi^{Dt}(p_0, p_i, c) + w_i^{Cs} \cdot m(p_i, id, c))]. \end{aligned} \tag{7}$$

By defining

$$\begin{aligned} R_0(p_0, id, c) = &\lambda^{Dt} w_0^{Dt} \cdot \phi^{Dt}(I, p_0, c) + \lambda^{Sg} w^{Sg} \cdot \phi^{Sg}(I, id) \\ &+ \lambda^{Ct} w^{Ct} \cdot \phi^{Ct}(I, id, p_0) + w_0^{Cs} \cdot m(p_0, id, c) \\ R_i(p_i, id, c) = &\lambda^{Dt} w_i^{Dt} \cdot \phi^{Dt}(I, p_i, c) + w_i^{Cs} \cdot m(p_i, id, c), \end{aligned}$$

the Eqn. (7) can be written compactly as:

$$S(p_0,c)=\max_{id}[R_0(p_0,id,c)+\max_{p_1,\cdots,p_m}\sum_{i=1}^{m}(R_i(p_i,id,c)+\lambda^{Dt}w_{i,def}^{Dt}\cdot\phi^{Dt}(p_0,p_i,c))]. \quad (8)$$

With fixed segment index $id$, this scoring function is similar to that of DPM and can thus be passed to an off-the-shelf DPM solver. Hence, the inference algorithm works as follows: First, we compute $R_0(p_0,id,c)$ for each root filter position $p_0$ and segment index $id$. Then, we prune the object hypotheses based on the score of $R_0$ without sacrificing the overall recall rate (validated on the validation set). For each retained segment hypothesis, we further run the full model (7) locally with the dynamic programming approach as in [19]. Finally, we compute the maximum over the mixture components to obtain the final score of the object hypothesis.

## 4.2 Learning

By defining the output variable $y=\{p_0,id\}$ and latent variable $h=\{p_1,\cdots,p_m,c\}$, the scoring function (1) can be rewritten as

$$S(I,y,h)=w\cdot\varPhi(I,y,h), \quad (9)$$

where $w$ is the concatenation of all model parameters ($w^{Dt}, w^{Sg}, w^{Ct}$ and $w^{Cs}$). $\varPhi(I,y,h)$ is the concatenation of all four components features weighted by their weights ($\lambda^{Dt}$, $\lambda^{Sg}$ and $\lambda^{Ct}$) with respect to the label $y$ and latent variable $h$.

We note that Eqn. (9) is linear in the model parameter $w$, thus this model can be effectively learned based on the latent structure SVM framework [40, 22]:

$$\min_{w}\frac{1}{2}||w||^2+C[\sum_{j=1}^{n}\max_{\hat{y},\hat{h}}(w\cdot\varPhi(x_j,\hat{y},\hat{h})+\varDelta(y_i,\hat{y},\hat{h}))-\sum_{j=1}^{n}\max_{h}(w\cdot\varPhi(x_i,y_i,h))], \quad (10)$$

where the loss function $\varDelta(y_i,\hat{y},\hat{h})$ is defined as the weighted sum of the Intersection over Union of the root filters and segment hypotheses (in current implementation, we simply use the average value of two IoUs).

The standard approach to solve the optimization problem (10) is the Concave-Convex Procedure (CCCP) [42, 40]. However, as the CCCP algorithm only guarantees to converge to a local minimum, we learn the model progressively to ensure a reasonable initialization. More specifically, we first train each component separately and jointly learn the overall model with Eqn (10).

For the object detection component, we follow the original training approach of DPM [19] except for the mixture initialization and part discovery. Aspect ratio based clustering is used in [19] for mixture initialization. However, such an approach may ignore the potential pose/view variance. Hence, we employ the idea of "subcategory mining" [15, 1, 14] by utilizing the additional segmentation annotation to ensure a more reliable shape mask for each component. Specifically, we resize all the cropped segmentation masks to the same height and $l_2$ normalizes all the resized masks. Then, the similarity between two normalized masks $a$ and $b$ is defined as the maximal value of the convolution response map of $a$ and $b$. Finally, the graph shift algorithm [30] is employed to discover the dense subgraphs, which correspond to the subcategories, as in [15]. The resulting subcategories are then used for mixture initialization. The original DPM approach [19] discovers the salient parts greedily by covering the high-energy region of the root HOG-template. Recently, [8] suggests that modifying this "saliency" measure by multiplying the HOG magnitude by the average segmentation mask for each component will lead to more semantic meaningful parts. Hence, we follow their approach by utilizing the modified 'saliency" measure for part discovery. For the consistency component, the pixel-wise mean of all segmentation masks for each component is utilized for initialization.

In the final joint learning stage, all model parameters ($w^{Dt}, w^{Sg}, w^{Ct}$ and $w^{Cs}$) in Eqn. (10) are jointly optimized. Thus, the relative importance of each component will be automatically tuned for each category.

### 4.3 Implementation Details

As discussed in Section 3.3 and 3.4, we employ the predicted scores of the basic-level classifiers as features for both the segmentation ($\phi^{Sg}(I, id)$ in Eqn. (4)) and context ($\phi^{Ct}(I, id, p_0)$ in Eqn. (6)) components to improve the efficiency of the UDS framework. For the segmentation component, we follow the second-order pooling approach [33] by utilizing the public available implementation provided by the author. 150 top-ranked object hypotheses are generated with the CPMC method for each image [7]. The concatenation of scores from support vector regressors of all categories is employed as the segmentation component feature for each hypothesis. For the context component, the dense SIFT [31] and color moment are extracted as low-level features. Both features are projected to 64 dimensions using PCA and the size of Gaussian Mixture Model in FV [9] is set to 64. The concatenation of resulting FVs in all regions is then trained with the LibLinear library [18] in a similar manner with [13]. Finally, the confidence scores of classifiers for all categories are utilized as the context component features.

For the shape-guided DPM, the number of subcategories is automatically decided by the graph shift algorithm based on the expansion size, which is decided by cross-validation [30]. The resulting subcategory number for different object categories is generally from 4 to 8.

The weights $\lambda^{Dt}, \lambda^{Sg}$ and $\lambda^{Ct}$ in Eqn. (1) are set as 0.1, 0.2 and 0.2, respectively, based on cross-validation. In fact, the final accuracy is not very sensitive to the variation of these parameters, as our UDS framework can automatically learn $w$ to adjust the relative weights of different components.

## 5 Experiments

We extensively evaluate the proposed UDS framework on the challenging PASCAL Visual Object Challenge (VOC) datasets [17], which provide a common

**Table 1.** Proof-of-Concept experiments for object detection on VOC 2010 validation set.

| Method | plane | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | motor | person | plant | sheep | sofa | train | tv | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DPM | 43.6 | 51.1 | 4.4 | 3.4 | 21.7 | 57.4 | 40.4 | 17.0 | 16.4 | 15.3 | 10.2 | 11.1 | 37.2 | 39.1 | 40.4 | 5.2 | 27.4 | 18.9 | 39.7 | 37.1 | 26.9 |
| S-DPM | 48.2 | 52.7 | 4.9 | 5.7 | 25.3 | 60.6 | 40.8 | 21.6 | **16.6** | 16.3 | 17.0 | 12.5 | 40.5 | 38.8 | **41.3** | 6.9 | 32.5 | 23.2 | 44.3 | 40.8 | 29.5 |
| S-DPM+Sg | 57.6 | 55.4 | 22.6 | 15.8 | 27.9 | **64.3** | 45.8 | 54.8 | 10.7 | 26.9 | 21.9 | 35.2 | 48.2 | 49.8 | 38.8 | 13.3 | 36.3 | 32.5 | 49.0 | 45.3 | 37.6 |
| S-DPM+Sg+Ct | **59.2** | **56.7** | **22.8** | **16.4** | **28.9** | 63.7 | **46.6** | **56.2** | 15.6 | **29.1** | **25.1** | **36.9** | **49.5** | **50.7** | 39.3 | **14.4** | **38.2** | **36.1** | **49.2** | **46.2** | **39.0** |

**Table 2.** Proof-of-Concept experiments for semantic segmentation on VOC 2010 validation set.

| Method | b/g | plane | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | motor | person | plant | sheep | sofa | train | tv | avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| O$_2$P | 83.2 | 70.0 | 22.0 | 43.8 | 39.6 | 40.3 | 60.3 | 64.9 | 55.7 | 13.2 | 37.1 | 20.2 | 42.5 | **37.3** | 47.1 | 50.5 | 31.9 | 51.5 | 27.2 | 58.6 | 50.6 | 45.1 |
| S-DPM+ Sg | 82.5 | 74.2 | 20.5 | 45.0 | 42.7 | 38.4 | 65.1 | 66.9 | 55.8 | 16.1 | 37.3 | 23.3 | 41.3 | 34.7 | 49.6 | 49.5 | 34.1 | 54.6 | 33.4 | 63.7 | 53.5 | 46.8 |
| S-DPM+Sg+Ct | **83.2** | **74.9** | **22.9** | **45.7** | **43.4** | **40.6** | **66.2** | **68.1** | **56.4** | **16.8** | **39.8** | **24.0** | **44.2** | 36.3 | **49.9** | **50.9** | **34.4** | **56.7** | **34.1** | **64.8** | **54.4** | **48.0** |

evaluation platform for both object detection and semantic segmentation. These datasets are extremely challenging since the images are crawled from the real-world photo sharing website and the objects contained vary significantly in size, pose, view point and appearance. The datasets contain 20 object classes and are divided into “train”, “val” and “test” subsets. We follow the standard PASCAL protocol by employing Average Precision (AP) and Intersection over Union (IoU) as evaluation metric for object detection and semantic segmentation, respectively.

In the following section, we first conduct multiple Proof-of-Concept experiments on the validation set to assess the relative importance of each individual component. Then, we evaluate the optimal configuration of the proposed framework on the test set to compare with the state-of-the-art performance ever reported for both object detection and semantic segmentation tasks.

### 5.1 Proof-of-Concept Experiments

In this subsection, we evaluate the relative importance of individual components in our framework on VOC 2012 “train/val” datasets (i.e. “train” set for training and “val” set for test) with the extra segmentation annotation from [24] for proof of concept and ease of parameter tuning.

Table 1 and 2 show the detailed object detection and semantic segmentation results, respectively. It can be concluded from the tables that:

- Shape-guided subcategory mining does improve the detection performance. By better capturing the pose/viewpoint variance and adaptively deciding the number of subcategories, shape-guided DPM (S-DPM) can provide more reliable shape masks for our UDS framework.
- Object detection and semantic segmentation techniques are complementary. Performing two tasks jointly will boost the performance of each other. As shown in Table 1, the joint approach (S-DPM+Sg) significantly outperforms

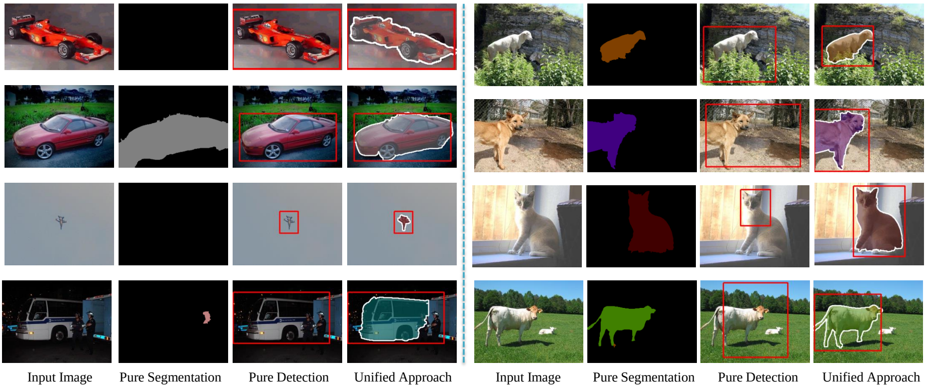


**Fig. 4.** More exemplar results on VOC 2012 from the proposed UDS framework and baseline methods (DPM [19] for detection and $O_2P$ [6] for segmentation).

the detection baseline (S-DPM) by 8.1%. In fact, the DPM based detector mainly captures the shape cues. Hence, it may locate rigid parts only and thus leads to localization error. On the contrary, the underlying segmentation component mainly relies on the appearance cues and thus can help to rectify the bounding box position, especially for the objects with homogeneous appearances. Table 2 demonstrates that the joint approach (S-DPM+Sg) also outperforms the segmentation baseline ($O_2P$). For objects in the cluttered background, shape based detectors can provide valuable information to assist in selecting better segment hypotheses. More examples to illustrate the complementarity of the two tasks are shown in Figure 4.

- The context component can further improve the performance for both tasks. By employing both the local and global context cues, the full model (S-DPM+Sg+Ct) can better distinguish ambiguous objects and thus yields the best performance.

### 5.2 Comparison with State-of-the-arts

In this subsection, we evaluate our UDS framework on the Pasval VOC test set to have a direct comparison with the state-of-the-arts. Though our framework can perform joint detection and segmentation, these two tasks are usually evaluated using different image sets. Hence, we slightly tweak the training process to allow the direct comparison with previous methods. Specifically, for the detection task, we train the model on the VOC 2010 "main-trainval" set, as many leading methods [20, 11] only reported their results on this dataset. For the segmentation task, we perform the experiments on the union of the VOC 2012 "main" and "seg" sets. The extra segmentation annotation from [24] are used for both tasks. We omit the results of VOC 2010 segmentation and VOC 2012 detection due to space limitation.

**Table 3.** Comparison of detection performance on VOC 2010 test set.

| Method | plane | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | motor | person | plant | sheep | sofa | train | tv | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DPM [19] | 48.2 | 52.2 | 14.8 | 13.8 | 28.7 | 53.2 | 44.9 | 26.0 | 18.4 | 24.4 | 13.7 | 23.1 | 45.8 | 50.5 | 43.7 | 9.8 | 31.1 | 21.5 | 44.4 | 35.7 | 32.2 |
| van de Sande *et al.* [36] | 56.2 | 42.4 | 15.3 | 12.6 | 21.8 | 49.3 | 36.8 | 46.1 | 12.9 | 32.1 | 30.0 | 36.5 | 43.5 | 52.9 | 32.9 | **15.3** | 41.1 | 31.8 | 47.0 | 44.8 | 35.1 |
| Gu *et al.* [23] | 53.7 | 42.9 | 18.1 | 16.5 | 23.5 | 48.1 | 42.1 | 45.4 | 6.7 | 23.4 | 27.7 | 35.2 | 40.7 | 49.0 | 32.0 | 11.6 | 34.6 | 28.7 | 43.3 | 39.2 | 33.1 |
| NLPR [17] | 53.3 | **55.3** | 19.2 | 21.0 | 30.0 | 54.4 | 46.7 | 41.2 | **20.0** | 31.5 | 20.7 | 30.3 | 48.6 | 55.3 | **46.5** | 10.2 | 34.4 | 26.5 | 50.3 | 40.3 | 36.8 |
| MITUCLA [43] | 54.2 | 48.5 | 15.7 | 19.2 | 29.2 | 55.5 | 43.5 | 41.7 | 16.9 | 28.5 | 26.7 | 30.9 | 48.3 | 55.0 | 41.7 | 9.7 | 35.8 | 30.8 | 47.2 | 40.8 | 36.0 |
| ContextSVM [34] | 53.1 | 52.7 | 18.1 | 13.5 | 30.7 | 53.9 | 43.5 | 40.3 | 17.7 | 31.9 | 28.0 | 29.5 | **52.9** | 56.6 | 44.2 | 12.6 | 36.2 | 28.7 | 50.5 | 40.7 | 36.8 |
| FV [11] | **65.9** | 50.1 | 23.7 | **24.1** | 20.4 | 52.6 | 47.1 | 50.9 | 13.2 | 32.8 | **31.8** | 41.4 | 43.9 | 55.3 | 29.8 | 14.1 | **41.7** | 35.6 | 46.7 | **46.9** | 38.4 |
| | Using Extra Semantic Segmentation Annotation From [24] | | | | | | | | | | | | | | | | | | | | |
| segDPM [20] | 58.7 | 51.4 | **25.3** | **24.1** | **33.8** | 52.5 | 49.2 | 48.8 | 11.7 | 30.4 | 21.6 | 37.7 | 46.0 | 53.1 | 46.0 | 13.1 | 35.7 | 29.4 | 52.5 | 41.8 | 38.1 |
| Ours:UDS | 60.1 | 54.3 | 23.9 | 22.9 | 31.8 | **57.0** | **51.1** | **54.8** | 17.6 | **35.7** | 26.7 | **42.8** | 51.2 | **58.0** | 41.7 | **15.3** | 37.8 | **39.8** | **54.9** | 45.6 | **41.2** |

**Object Detection:** The detailed comparison of the proposed framework with current leading approaches for object detection is presented in Table 3. The first two methods represent two different lines of approaches for object detection. DPM [19] employed shape based templates with the sliding window strategy while van de Sande *et al.* [36] utilized the appearance based model with the selective window strategy. Gu *et al.* [23] further extended DPM with a multiple component mechanism. Despite their theoretical interest, these methods only focus on the information within the windows and thus ignore the informative context cues, which leads to inferior results compared with other competitors. All other methods are obtained through the combinations of multiple techniques in order to obtain better performance.

From Table 3, it can be observed that our proposed UDS outperforms all the competitors in terms of mAP. The proposed UDS framework achieves the best performance in 9 out of the 20 categories with an mAP of 41.2%, which is 3.1% higher than that of the state-of-the-arts. With our unified approach, the advantages of both object detection and semantic segmentation techniques can be leveraged to improve the overall performance. In addition, it can be noted that our method can significantly improve the performance on the categories with homogeneous appearances, such as cats and dogs. For such categories, the underlying segmentation component can easily segment the objects out for rectifying the localization errors.

**Semantic Segmentation:** Table 4 shows the detailed comparison of the proposed framework with previous approaches on the VOC 2012 segmentation challenge. Based on the basic idea behind the methods, all the competing methods can be divided into two categories. The first category (O2P-CPMC-CSI, CMBR-O2P-CPMC-LIN, O2P-CPMC-FGT-SEGM and Yadollahpour) employs the hypotheses based segmentation. The difference among them mainly lies in the hypotheses generation procedure and ranking function design. Most of them provide the results with/without extra annotation from [24]. The other category (NUS-DET-SPR-GC-SP and Xia) estimates the semantic segmentation results based on the bounding boxes from object detection. Hence, these approaches heavily rely on the detector performance and need extra annotation for object detection.

**Table 4.** Comparison of segmentation performance on VOC 2012 test set.

| Method | b/g | plane | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | motor | person | plant | sheep | sofa | train | tv | avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| O2P-CPMC-CSI [16] | 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 [16] | 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 [16] | 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 |
| Yadollahpour *et al.* [38] | **85.7** | 62.7 | 25.6 | 46.9 | 43.0 | 54.8 | 58.4 | 58.6 | 55.6 | 14.6 | 47.5 | 31.2 | 44.7 | 51.0 | 60.9 | **53.5** | 36.6 | 50.9 | 30.1 | 50.2 | 46.8 | 48.1 |
| | Relying on Extra Object Detector | | | | | | | | | | | | | | | | | | | | | |
| NUS-DET-SPR-GC-SP [16] | 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 |
| Xia *et al.* [37] | 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 |
| | Using Extra Semantic Segmentation Annotation From [24] | | | | | | | | | | | | | | | | | | | | | |
| O2P-CPMC-CSI [16] | 85.0 | 63.6 | 26.8 | 45.6 | 41.7 | 47.1 | 54.3 | 58.6 | 55.1 | 14.5 | 49.0 | 30.9 | 46.1 | 52.6 | 58.2 | 53.4 | 32.0 | 44.5 | 34.6 | 45.3 | 43.1 | 46.8 |
| CMBR-O2P-CPMC-LIN [16] | 84.7 | 63.9 | 23.8 | 44.6 | 40.3 | 45.5 | 59.6 | 58.7 | 57.1 | 11.7 | 45.9 | 34.9 | 43.0 | 54.9 | 58.0 | 51.5 | 34.6 | 44.1 | 29.9 | 50.5 | 44.5 | 46.7 |
| O2P-CPMC-FGT-SEGM [16] | 85.2 | 63.4 | 27.3 | **56.1** | 37.7 | 47.2 | 57.9 | 59.3 | 55.0 | 11.5 | 50.8 | 30.5 | 45.0 | 58.4 | 57.4 | 48.6 | 34.6 | 53.3 | 32.4 | 47.6 | 39.2 | 47.5 |
| Ours:UDS | 85.2 | **67.0** | 24.5 | 47.2 | **45.0** | 47.9 | **65.3** | 60.6 | **58.5** | 15.5 | 50.8 | **37.4** | 45.8 | **59.9** | **62.0** | 52.7 | **40.8** | 48.2 | 36.8 | 53.1 | 45.6 | **50.0** |

The results in Table 4 demonstrate that the proposed UDS framework performs the best in 8 out of the 21 categories, achieving the best average performance of 50%. As discussed above, our unified approach can leverage the advantages of both object detection and semantic segmentation techniques. One main source of the improvement for semantic segmentation comes from the successful detection of objects in cluttered backgrounds. The bottom-up segmentation techniques may not be able to extract the accurate boundary of objects in cluttered backgrounds, which makes the following ranking problem very difficult. However, the template based detection mainly focuses on the object shape and thus is robust to the cluttered backgrounds to some extent. Hence, the proposed framework can significantly improve the semantic segmentation performance of rigid objects, such as aeroplane, bus and motorbike, as verified in Table 4.

## 6 Conclusions and Future Work

In this paper, we proposed a unified framework for joint object detection and semantic segmentation. Noticing the complementarity of current detection and segmentation approaches, we explicitly enforce the consistency between their outputs to leverage the advantages of both techniques. Both local and global context information are further integrated into the framework to better distinguish the ambiguous samples. All the information is aggregated at the end of the pipeline for decision making and thus hard decision is avoided to make at the early stage as in traditional pipelines. The relative importance of different components is automatically learned for each category to guarantee the overall performance. Extensive experimental results clearly demonstrated the proposed framework has achieved the state-of-the-art performance. In the future, we plan to integrate deep learning techniques into the current framework.

## 7 Acknowledgment

This research is supported by the Singapore National Research Foundation under its International Research Centre @Singapore Funding Initiative, Office of Naval Research ONR N00014-12-1-0883 and ONR MURI N000014-10-1-0933.

## References

1. Aghazadeh, O., Azizpour, H., Sullivan, J., Carlsson, S.: Mixture component identification and learning for visual recognition. In: ECCV (2012)
2. Arbeláez, P., Hariharan, B., Gu, C., Gupta, S., Bourdev, L., Malik, J.: Semantic segmentation using regions and parts. In: CVPR (2012)
3. Boix, X., Gonfaus, J.M., van de Weijer, J., Bagdanov, A.D., Serrat, J., Gonzàlez, J.: Harmony potentials. IJCV (2012)
4. Brox, T., Bourdev, L., Maji, S., Malik, J.: Object segmentation by alignment of poselet activations to image contours. In: CVPR (2011)
5. Brox, T., Bourdev, L., Maji, S., Malik, J.: Object segmentation by alignment of poselet activations to image contours. In: CVPR (2011)
6. Carreira, J., Caseiro, R., Batista, J., Sminchisescu, C.: Semantic segmentation with second-order pooling. In: ECCV (2012)
7. Carreira, J., Sminchisescu, C.: Cpmc: Automatic object segmentation using constrained parametric min-cuts. TPAMI (2012)
8. Chai, Y., Lempitsky, V., Zisserman, A.: Symbiotic segmentation and part localization for fine-grained categorization. ICCV (2013)
9. Chatfield, K., Lempitsky, V., Vedaldi, A.: The devil is in the details: an evaluation of recent feature encoding methods. In: BMVC (2011)
10. Chen, Q., Song, Z., Hua, Y., Huang, Z., Yan, S.: Hierarchical matching with side information for image classification. In: CVPR (2012)
11. Cinbis, R.G., Verbeek, J., Schmid, C., et al.: Segmentation driven object detection with fisher vectors. In: ICCV (2013)
12. Dai, Q., Hoiem, D.: Learning to localize detected objects. In: CVPR (2012)
13. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
14. Divvala, S.K., Efros, A.A., Hebert, M.: How important are ”deformable parts” in the deformable parts model? In: ECCV Workshops (2012)
15. Dong, J., Xia, W., Chen, Q., Feng, J., Huang, Z., Yan, S.: Subcategory-aware object classification. In: CVPR (2013)
16. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results
17. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision 88(2), 303–338 (Jun 2010)
18. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. JMLR (2008)
19. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object Detection with Discriminatively Trained Part-Based Models. TPAMI (2010)
20. Fidler, S., Mottaghi, R., Yuille, A., Urtasun, R.: Bottom-up segmentation for top-down detection. In: CVPR (2013)
21. Florent Perronnin, J.S., Mensink, T.: Improving the Fisher Kernel for Large-Scale Image Classification. In: ECCV (2010)
22. Girshick, R.B., Felzenszwalb, P., Mcallester, D.: Object detection with grammar models. In: NIPS (2011)
23. Gu, C., Arbeláez, P.A., Lin, Y., Yu, K., Malik, J.: Multi-component models for object detection. In: ECCV (2012)
24. Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: ICCV (2011)

25. Hoiem, D., Chodpathumwan, Y., Dai, Q.: Diagnosing error in object detectors. In: Computer Vision–ECCV 2012, pp. 340–353. Springer (2012)
26. Kumar, M.P., Ton, P., Zisserman, A.: Obj cut. In: CVPR (2005)
27. Ladickỳ, L., Sturgess, P., Alahari, K., Russell, C., Torr, P.H.: What, where and how many? combining object detectors and crfs. In: ECCV (2010)
28. Lempitsky, V., Kohli, P., Rother, C., Sharp, T.: Image segmentation with a bounding box prior. In: Computer Vision, 2009 IEEE 12th International Conference on (2009)
29. Li, C., Parikh, D., Chen, T.: Extracting adaptive contextual cues from unlabeled regions. In: ICCV (2011)
30. Liu, H., Yan, S.: Robust graph mode seeking by graph shift. In: ICML (2010)
31. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV (2004)
32. Parkhi, O.M., Vedaldi, A., Jawahar, C., Zisserman, A.: The truth about cats and dogs. In: ICCV (2011)
33. Russakovsky, O., Lin, Y., Yu, K., Fei-Fei, L.: Object-centric spatial pooling for image classification. In: ECCV (2012)
34. Song, Z., Chen, Q., Huang, Z., Hua, Y., Yan, S.: Contextualizing object detection and classification. In: CVPR (2011)
35. Tighe, J., Lazebnik, S.: Finding things: Image parsing with regions and per-exemplar detectors. In: CVPR (2013)
36. Uijlings, J., van de Sande, K., Gevers, T., Smeulders, A.: Selective search for object recognition. IJCV (2013)
37. Xia, W., Domokos, C., Dong, J., Cheong, L.F., Yan, S.: Semantic segmentation without annotating segments (2013)
38. Yadollahpour, P., Batra, D., Shakhnarovich, G.: Discriminative re-ranking of diverse segmentations. In: CVPR (2013)
39. Yang, Y., Hallman, S., Ramanan, D., Fowlkes, C.C.: Layered object models for image segmentation. PAMI (2012)
40. Yu, C.N.J., Joachims, T.: Learning structural svms with latent variables. In: ICML (2009)
41. Yuen, J., Zitnick, C.L., Liu, C., Torralba, A.: A framework for encoding object-level image priors. Tech. rep., Microsoft Research Technical Report
42. Yuille, A.L., Rangarajan, A.: The concave-convex procedure. Neural Computation (2003)
43. Zhu, L., Chen, Y., Yuille, A.L., Freeman, W.T.: Latent hierarchical structural learning for object detection. In: CVPR (2010)