Unconstrained Salient Object Detection via Proposal Subset Optimization

Jianming Zhang[1] Stan Sclaroff[1] Zhe Lin[2] Xiaohui Shen[2] Brian Price[2] Radomír Měch[2]
[1]Boston University [2]Adobe Research

Abstract

We aim at detecting salient objects in unconstrained images. In unconstrained images, the number of salient objects (if any) varies from image to image, and is not given. We present a salient object detection system that directly outputs a compact set of detection windows, if any, for an input image. Our system leverages a Convolutional-Neural-Network model to generate location proposals of salient objects. Location proposals tend to be highly overlapping and noisy. Based on the Maximum a Posteriori principle, we propose a novel subset optimization framework to generate a compact set of detection windows out of noisy proposals. In experiments, we show that our subset optimization formulation greatly enhances the performance of our system, and our system attains 16-34% relative improvement in Average Precision compared with the state-of-the-art on three challenging salient object datasets.

1. Introduction

In this paper, we aim at detecting generic salient objects in unconstrained images, which may contain multiple salient objects or no salient object. Solving this problem entails generating a compact set of detection windows that matches the number and the locations of salient objects. To be more specific, a satisfying solution to this problem should answer the following questions:

1. (Existence) Is there any salient object in the image?
2. (Localization) Where is each salient object, if any?

These two questions are important not only in a theoretic aspect, but also in an applicative aspect. First of all, a compact and clean set of detection windows can significantly reduce the computational cost of the subsequent process (*e.g.* object recognition) applied on each detection window [22, 36]. Furthermore, individuating each salient object (or reporting that no salient object is present) can critically alleviate the ambiguity in the weakly supervised or unsupervised learning scenario [10, 26, 55], where object appearance models are to be learned with no instance level annotation.

Figure 1: Our system outputs a compact set of detection windows (shown in the bottom row) that localize each salient object in an image. Note that for the input image in the right column, where no dominant object exists, our system does not output any detection window.

However, many previous methods [1, 11, 41, 30, 25, 6, 54] only solve the task of foreground segmentation, i.e. generating a dense foreground mask (saliency map). These methods do not individuate each object. Moreover, they do not directly answer the question of Existence. In this paper, we will use the term *salient region detection* when referring to these methods, so as to distinguish from the *salient object detection* task solved by our approach, which includes individuating each of the salient objects, if there are any, in a given input image.

Some methods generate a ranked list of bounding box candidates for salient objects [21, 43, 52], but they lack an effective way to fully answer the questions of Existence and Localization. In practice, they just produce a fixed number of location proposals, without specifying the exact set of detection windows. Other salient object detection methods simplify the detection task by assuming the existence of one and only one salient object [48, 45, 32]. This overly strong assumption limits their usage on unconstrained images.

In contrast to previous works, we present a salient object detection system that directly outputs a compact set of detections windows for an unconstrained image. Some example outputs of our system are shown in Fig. 1.

Our system leverages the high expressiveness of a Convolutional Neural Network (CNN) model to generate a set of scored salient object proposals for an image. Inspired by

the attention-based mechanisms of [27, 4, 35], we propose an Adaptive Region Sampling method to make our CNN model "look closer" at promising images regions, which substantially increases the detection rate. The obtained proposals are then filtered to produce a compact detection set.

A key difference between salient object detection and object class detection is that saliency greatly depends on the surrounding context. Therefore, the salient object proposal scores estimated on local image regions can be inconsistent with the ones estimated on the global scale. This intrinsic property of saliency detection makes our proposal filtering process challenging. We find that using the greedy Non-maximum Suppression (NMS) method often leads to sub-optimal performance in our task. To attack this problem, we propose a subset optimization formulation based on the *maximum a posteriori* (MAP) principle, which jointly optimizes the number and the locations of detection windows. The effectiveness of our optimization formulation is validated on various benchmark datasets, where our formulation attains about 12% relative improvement in Average Precision (AP) over the NMS approach.

In experiments, we demonstrate the superior performance of our system on three benchmark datasets: MSRA [29], DUT-O [51] and MSO [53]. In particular, the MSO dataset contains a large number of background/cluttered images that do not contain any dominant object. Our system can effectively handle such unconstrained images, and attains about 16-34% relative improvement in AP over previous methods on these datasets.

To summarize, the main contributions of this work are:

- A salient object detection system that outputs compact detection windows for *unconstrained* images,
- A novel MAP-based subset optimization formulation for filtering bounding box proposals,
- Significant improvement over the state-of-the-art methods on three challenging benchmark datasets.

2. Related Work

We review some previous works related to our task.

Salient region detection. Salient region detection aims at generating a dense foreground mask (saliency map) that separates salient objects from the background of an image [1, 11, 41, 50, 25]. Some methods allow extraction of multiple salient objects [33, 28]. However, these methods do not individuate each object.

Salient object localization. Given a saliency map, some methods find the best detection window based on heuristics [29, 48, 45, 32]. Various segmentation techniques are also used to generate binary foreground masks to facilitate object localization [29, 34, 23, 31]. A learning-based regression approach is proposed in [49] to predict a bounding box for an image. Most of these methods critically rely on the assumption that there is only one salient object in an image. In [29, 31], it is demonstrated that segmentation-based methods can localize multiple objects in some cases by tweaking certain parts in their formulation, but they lack a principled way to handle general scenarios.

Predicting the existence of salient objects. Existing salient object/region detection methods tend to produce undesirable results on images that contain no dominant salient object [49, 6]. In [49, 40], a binary classifier is trained to detect the existence of salient objects before object localization. In [53], a salient object subitizing model is proposed to suppress the detections on background images that contain no salient object. While all these methods use a separately trained background image detector, we provide a unified solution to the problems of Existence and Localization through our subset optimization formulation.

Object proposal generation. Object proposal methods [2, 9, 56, 47, 3, 12] usually generate hundreds or thousands of proposal windows in order to yield a high recall rate. While they can lead to substantial speed-ups over sliding window approaches for object detection, these proposal methods are not optimized for localizing salient objects. Some methods [43, 21] generate a ranked list of proposals for salient objects in an image, and can yield accurate localization using only the top few proposals. However, these methods do not aim to produce a compact set of detection windows that exactly match the ground-truth objects.

Bounding box filtering and NMS. Object detection and proposal methods often produce severely overlapping windows that correspond to a single object. To alleviate this problem, greedy Non-Maximum Suppression (NMS) is widely used due to its simplicity [13, 20, 2, 21]. Several limitations of greedy NMS are observed and addressed by [37, 5, 14, 38]. In [5], an improved NMS method is proposed for Hough transform based object detectors. Desai *et al*. [14] use a unified framework to model NMS and object class co-occurrence via Context Cueing. These methods are designed for a particular detection framework, which requires either part-based models or object category information. In [37], Affinity Propagation Clustering is used for bounding box filtering. This method achieves more accurate bounding box localization, but slightly compromises Average Precision (AP). In [38], Quadratic Binary Optimization is proposed to recover missing detections caused by greedy NMS. Unlike [37, 38], our subset optimization formulation aims to handle highly noisy proposal scores, where greedy NMS often leads to a poor detection precision rate.

3. A Salient Object Detection Framework

Our salient object detection framework comprises two steps. It first generates a set of scored location proposals using a CNN model. It then produces a compact set of detections out of the location proposals using a subset opti-

mization formulation. We first present the subset optimization formulation, as it is independent of the implementation of our proposal generation model, and can be useful beyond the scope of salient object detection.

Given a set of scored proposal windows, our formulation aims to extract a compact set of detection windows based on the following observations.

I. The scores of location proposals can be noisy, so it is often suboptimal to consider each proposal's score independently. Therefore, we jointly consider the scores and the spatial proximity of all proposal windows for more robust localization.

II. Severely overlapping windows often correspond to the same object. On the other hand, salient objects can also overlap each other to varying extents. We address these issues by *softly* penalizing overlaps between detection windows in our optimization formulation.

III. At the same time, we favor a compact set of detections that explains the observations, as salient objects are distinctive and rare in nature [16].

3.1. MAP-based Proposal Subset Optimization

Given an image I, a set of location proposals $\mathbf{B} = \{b_i : i = 1 \dots n\}$ and a proposal scoring function $\mathcal{S}$, we want to output a set of detection windows $\mathbf{O}$, which is a subset of $\mathbf{B}$. We assume each proposal b_i is a bounding box, with a score $s_i \triangleq \mathcal{S}(b_i, I)$. Given $\mathbf{B}$, the output set $\mathbf{O}$ can be represented as a binary indicator vector $(O_i)_{i=1}^n$, where $O_i = 1$ iff b_i is selected as an output.

The high-level idea of our formulation is to perform three tasks altogether: 1) group location proposals into clusters, 2) select an exemplar window from each cluster as an output detection, and 3) determine the number of clusters. To do so, we introduce an auxiliary variable $\mathbf{X} = (x_i)_{i=1}^n$. $\mathbf{X}$ represents the group membership for each proposal in $\mathbf{B}$, where $x_i = j$ if b_i belongs to a cluster represented by b_j. We also allow $x_i = 0$ if b_i does not belong to any cluster. Alternately, we can think that b_i belongs to the background. We would like to find the MAP solution w.r.t. the joint distribution $P(\mathbf{O}, \mathbf{X}|I; \mathbf{B}, \mathcal{S})$. In what follows, we omit the parameters $\mathbf{B}$ and $\mathcal{S}$ for brevity, as they are fixed for an image. According to Bayes' Rule, the joint distribution under consideration can be decomposed as

$$P(\mathbf{O}, \mathbf{X}|I) = \frac{P(I|\mathbf{O}, \mathbf{X})P(\mathbf{O}, \mathbf{X})}{P(I)}. \tag{1}$$

For the likelihood term $P(I|\mathbf{O}, \mathbf{X})$, we assume that $\mathbf{O}$ is conditionally independent of I given $\mathbf{X}$. Thus,

$$\begin{aligned} P(I|\mathbf{O}, \mathbf{X}) &= P(I|\mathbf{X}) \\ &= \frac{P(\mathbf{X}|I)P(I)}{P(\mathbf{X})}. \end{aligned} \tag{2}$$

The conditional independence assumption is natural, as the detection set $\mathbf{O}$ can be directly induced by the group membership vector $\mathbf{X}$. In other words, representative windows indicated by $\mathbf{X}$ should be regarded as detections windows. This leads to the following constraint on $\mathbf{X}$ and $\mathbf{O}$:

Constraint 1. If $\exists x_i$ s.t. $x_i = j,\ j \neq 0$, then $b_j \in \mathbf{O}$.

To comply with this constraint, the prior term $P(\mathbf{O}, \mathbf{X})$ takes the following form:

$$P(\mathbf{O}, \mathbf{X}) = Z_1 P(\mathbf{X}) L(\mathbf{O}) \mathcal{C}(\mathbf{O}, \mathbf{X}), \tag{3}$$

where $\mathcal{C}(\mathbf{O}, \mathbf{X})$ is a constraint compliance indicator function, which takes 1 if Constraint 1 is met, and 0 otherwise. Z_1 is a normalization constant that makes $P(\mathbf{O}, \mathbf{X})$ a valid probability mass function. The term $L(\mathbf{O})$ encodes prior information about the detection windows. The definition of $P(\mathbf{O}, \mathbf{X})$ assumes the minimum dependency between $\mathbf{O}$ and $\mathbf{X}$ when Constraint 1 is met.

Substituting Eq. 2 and 3 into the RHS of Eq. 1, we have

$$P(\mathbf{O}, \mathbf{X}|I) \propto P(\mathbf{X}|I) L(\mathbf{O}) \mathcal{C}(\mathbf{O}, \mathbf{X}). \tag{4}$$

Note that both $P(I)$ and $P(\mathbf{X})$ are cancelled out, and the constant Z_1 is omitted.

3.2. Formulation Details

We now provide details for each term in Eq. 4, and show the connections with the observations we made.

Assuming that the x_i are independent of each other given I, we compute $P(\mathbf{X}|I)$ as follows:

$$P(\mathbf{X}|I) = \prod_{i=1}^{n} P(x_i|I), \tag{5}$$

where

$$P(x_i = j|I) = \begin{cases} Z_2^i \lambda & \text{if } j = 0; \\ Z_2^i \mathcal{K}(b_i, b_j) s_i & \text{otherwise.} \end{cases} \tag{6}$$

Here Z_2^i is a normalization constant such that $\sum_{j=0}^{n} P(x_i = j|I) = 1$. $\mathcal{K}(b_i, b_j)$ is a function that measures the spatial proximity between b_i and b_j. We use window Intersection Over Union (IOU) [37, 18] as $\mathcal{K}$. The parameter λ controls the probability that a proposal window belongs to the background. The formulation of $P(\mathbf{X}|I)$ favors representative windows that have strong overlap with many confident proposals. By jointly considering the scores and the spatial proximity of all proposal windows, our formulation is robust to individual noisy proposals. This addresses Observation **I**.

Prior information about detection windows is encoded in $L(\mathbf{O})$, which is formulated as

$$L(\mathbf{O}) = L_1(\mathbf{O}) L_2(|\mathbf{O}|), \tag{7}$$

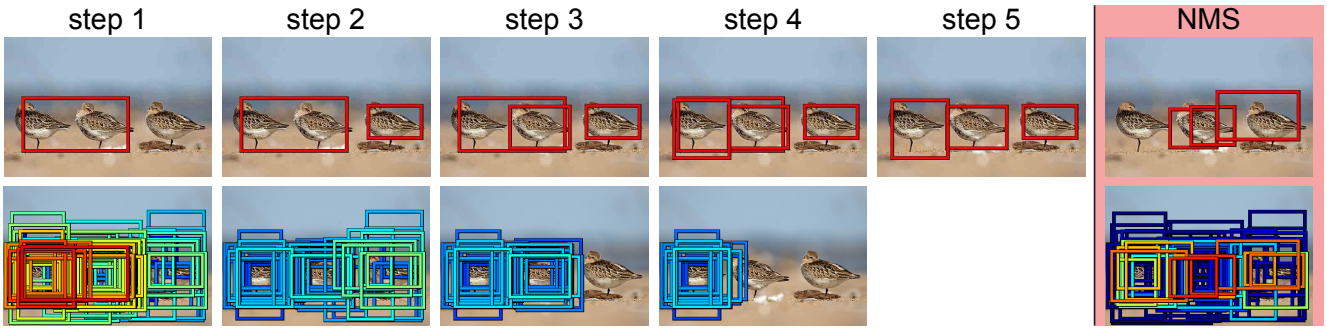

Figure 2: In column 1-5, we show step-by-step window selection results of our greedy algorithm. In the incrementing pass (step 1-4), windows are selected based on their marginal gains w.r.t. Eq. 11. The window proposals of positive marginal gains are shown in the bottom row for each step. Warmer colors indicate higher marginal gains. The final step (step 5) removes the first selected window in the decrementing pass, because our formulation favors a small number of detection windows with small inter-window overlap. To contrast our method with greedy NMS, we show the top 3 output windows after greedy NMS using an IOU threshold of 0.4 (top). The scored proposals are shown in the bottom row of the figure.

where

$$L_1(\mathbf{O}) = \prod_{i,j:i\neq j} \exp\left(-\frac{\gamma}{2} O_i O_j \mathcal{K}(b_i, b_j)\right). \quad (8)$$

$L_1(\mathbf{O})$ addresses Observation **II** by penalizing overlapping detection windows. Parameter γ controls the penalty level.

$L_2(|\mathbf{O}|)$ represents the prior belief about the number of salient objects. According to Observation **III**, we favor a small set of output windows that explains the observation. Therefore, $L_2(.)$ is defined as

$$L_2(N) = \exp(-\phi N), \quad (9)$$

where ϕ controls the strength of this prior belief.

Our MAP-based formulation answers the question of Localization by jointly optimizing the number and the locations of the detection windows, and it also naturally addresses the question of Existence, as the number of detections tends to be zero if no strong evidence of salient objects is found (Eq. 9). Note that $L(\mathbf{O})$ can also be straightforwardly modified to encode other priors regarding the number or the spatial constraints of detection windows.

3.3. Optimization

Taking the log of Eq. 4, we obtain our objective function:

$$f(\mathbf{O}, \mathbf{X}) = \sum_{i=1}^{n} w_i(x_i) - \phi|\mathbf{O}| - \frac{\gamma}{2} \sum_{i,j\in\widetilde{\mathbf{O}}:i\neq j} \mathcal{K}_{ij}, \quad (10)$$

where $w_i(x_i = j) \triangleq \log P(x_i = j|I)$ and $\mathcal{K}_{ij}$ is shorthand for $\mathcal{K}(b_i, b_j)$. $\widetilde{\mathbf{O}}$ denotes the index set corresponding to the selected windows in $\mathbf{O}$. We omit $\log \mathcal{C}(\mathbf{O}, \mathbf{X})$ in Eq. 10, as we now explicitly consider Constraint 1.

Since we are interested in finding the optimal detection set $\mathbf{O}^*$, we can first maximize over $\mathbf{X}$ and define our optimization problem as

$$\mathbf{O}^* = \arg\max_{\mathbf{O}} \left(\max_{\mathbf{X}} f(\mathbf{O}, \mathbf{X})\right), \quad (11)$$

which is subject to Constraint 1. Given $\mathbf{O}$ is fixed, the subproblem of maximizing $f(\mathbf{O}, \mathbf{X})$ over $\mathbf{X}$ is straightforward:

$$\mathbf{X}^*(\mathbf{O}) = \arg\max_{\mathbf{X}} f(\mathbf{O}, \mathbf{X}) = \sum_{i=1}^{n} \max_{x_i \in \widetilde{\mathbf{O}} \cup \{0\}} w_i(x_i). \quad (12)$$

Let $h(\mathbf{O}) \triangleq f(\mathbf{O}, \mathbf{X}^*(\mathbf{O}))$, then Eq. 11 is equal to an unconstrained maximization problem of the set function $h(\mathbf{O})$, as Constraint 1 is already encoded in $\mathbf{X}^*(\mathbf{O})$.

The set function $h(\mathbf{O})$ is submodular (see proof in our supplementary material) and the maximization problem is NP-hard [19]. We use a simple greedy algorithm to solve our problem. Our greedy algorithm starts from an empty solution set. It alternates between an incrementing pass (Alg. 1) and a decrementing pass (Alg. 2) until a local minimum is reached. The incrementing (decrementing) pass adds (removes) the element with maximal marginal gain to (from) the solution set until no more elements can be added (removed) to improve the objective function. Convergence is guaranteed, as $h(\mathbf{O})$ is upper-bounded and each step of our algorithm increases $h(\mathbf{O})$. An example of the optimization process is shown in Fig. 2.

In practice, we find that our greedy algorithm usually converges within two passes, and it provides reasonable solutions. Some theoretic approximation analyses for unconstrained submodular maximization [7, 19] may shed light on the good performance of our greedy algorithm.

3.4. Salient Object Proposal Generation by CNN

We present a CNN-based approach to generate scored window proposals $\{(b_i, s_i)\}_{i=1}^{n}$ for salient objects. Inspired

Alg. 1 `IncrementPass`($\mathbf{O}$)

$\mathbf{V} \leftarrow \mathbf{B} \setminus \mathbf{O}$
while $\mathbf{V} \neq \varnothing$ **do**
  $b^* \leftarrow \arg\max_{b \in \mathbf{V}} h(\mathbf{O} \cup \{b\})$
  if $h(\mathbf{O} \cup \{b^*\}) > h(\mathbf{O})$ **then**
    $\mathbf{O} \leftarrow \mathbf{O} \cup \{b^*\}$
    $\mathbf{V} \leftarrow \mathbf{V} \setminus \{b^*\}$
  else
    return

Alg. 2 `DecrementPass`($\mathbf{O}$)

while $\mathbf{O} \neq \varnothing$ **do**
  $b^* \leftarrow \arg\max_{b \in \mathbf{O}} h(\mathbf{O} \setminus \{b\})$
  if $h(\mathbf{O} \setminus \{b^*\}) > h(\mathbf{O})$ **then**
    $\mathbf{O} \leftarrow \mathbf{O} \setminus \{b^*\}$
  else
    return

by [17, 46], we train a CNN model to produce a fixed number of scored window proposals. As our CNN model takes the whole image as input, it is able to capture context information for localizing salient objects. Our CNN model predicts scores for a predefined set of exemplar windows. Furthermore, an Adaptive Region Sampling method is proposed to significantly enhance the detection rate of our CNN proposal model.

Generating exemplar windows. Given a training set with ground-truth bounding boxes, we transform the coordinates of each bounding box to a normalized coordinate space, *i.e.* $(x, y) \rightarrow (\frac{x}{W}, \frac{y}{H})$, where W and H represents the width and height of the given image. Each bounding box is represented by a 4D vector composed of the normalized coordinates of its upper-left and bottom-right corners. Then we obtain K exemplar windows via K-means clustering in this 4D space. In our implementation, we set $K = 100$.

Adaptive region sampling. The 100 exemplar windows only provide a coarse sampling of location proposals. To address this problem, the authors of [17, 46] propose to augment the proposal set by running the proposal generation method on uniformly sampled regions of an image. We find this uniformly sampling inefficient for salient object detection, and sometimes it even worsens the performance in our task (see Sec. 4).

Instead, we propose an adaptive region sampling method, which is in a sense related to the attention mechanism used in [27, 4, 35]. After proposal generation on the whole image, our model takes a closer glimpse at those important regions indicated by the global prediction. To do so, we choose the top M windows generated by our CNN model for the whole image, and extract the corresponding sub-images after expanding the size of each window by 2X. We then apply our CNN model on each of these sub-images to augment our proposal set. In our implementation, we set $M = 5$, and only retain the top 10 proposals from each sub-image. This substantially speeds up the subsequent optimization process without sacrificing the performance.

The downside of the this adaptive region sampling is that it may introduce more noise into the proposal set, because the context of the sub-images can be very different from the whole image. This makes the subsequent bounding box filtering task more challenging.

CNN model architecture and training. We use the VGG16 model architecture [42], and replace its fc8 layer with a 100-D linear layer followed by a sigmoid layer. Let $(c_i)_{i=1}^K$ denote the output of our CNN model. Logistic loss $\sum_i -y_i \log c_i - (1 - y_i) \log(1 - c_i)$ is used to train our model, where the binary label $y_i = 1$ iff the i-th exemplar window is the nearest to a ground-truth bounding box in the 4D normalized coordinate space.

To train our CNN model, we use about 5500 images from the training split of the Salient Object Subitizing (SOS) dataset [53]. The SOS dataset comprises unconstrained images with varying numbers of salient objects. In particular, the SOS dataset has over 1000 background/cluttered images that contain no salient objects, as judged by human annotators. By including background images in the training set, our model is expected to suppress the detections on this type of images. As the SOS dataset only has annotations about the number of salient objects in an image, we manually annotated object bounding boxes according to the number of salient objects given for each image. We excluded a few images that we found ambiguous to annotate.

We set aside 1/5 of the SOS training images for validation purpose. We first fine-tune the pre-trained VGG16 model on the ILSVRC-2014 object detection dataset [39] using the provided bounding box annotations, and then fine-tune it using the SOS training set. We find this two-stage fine-tuning gives lower validation errors than only fine-tuning on the SOS training set. Training details are included in our supplementary material due to limited space.

Our full system and the bounding box annotations of the SOS training set are available on our project website[1].

4. Experiments

Evaluation Metrics. Following [21, 43], we use the PASCAL evaluation protocol [18] to evaluate salient object detection performance. A detection window is judged as correct if it overlaps with a ground-truth window by more than half of their union. We do not allow multiple detections for one object, which is different from the setting of [21]. Precision is computed as the percentage of correct predictions, and Recall is the percentage of detected ground-truth objects. We evaluate each method by 1) Precision-Recall (PR) curves, which are generated by varying a parameter for each method (see below), and 2) Average Precision (AP),

[1] http://www.cs.bu.edu/groups/ivc/SOD/

which is computed by averaging precisions on an interpolated PR curve at regular intervals (see [18] for details).

Precision-Recall Tradeoff. As our formulation does not generate scores for the detection windows, we cannot control the PR tradeoff by varying a score threshold. Here we provide a straightforward way to choose an operating point of our system. By varying the three parameters in our formulation, λ, γ and ϕ, we find that our system is not very sensitive to ϕ in Eq. 9, but responds actively to changes in λ and γ. λ controls the probability of a proposal window belonging to the background (Eq. 6), and γ controls the penalty for overlapping windows (Eq. 8). Thus, lowering either λ or γ increases the recall. We couple λ and γ by setting $\gamma = \alpha\lambda$, and fix ϕ and α in our system. In this way, the PR curve can be generated by varying λ. The parameters ϕ and α are optimized by grid search on the SOS training split. We fix ϕ at 1.2 and α at 10 for all experiments.

Compared Methods. Traditional salient region detection methods [1, 11, 41, 50, 25] cannot be fairly evaluated in our task, as they only generate saliency maps without individuating each object. Therefore, we mainly compare our method with two state-of-the-art methods, SC [21] and LBI [43], both of which output detection windows for salient objects. We also evaluate a recent CNN-based object proposal model, MultiBox (MBox) [17, 46], which is closely related to our salient object proposal method. MBox generates 800 proposal windows, and it is optimized to localize objects of certain categories of interest (*e.g.* 20 object classes in PASCAL VOC [18]), regardless whether they are salient or not.

These compared methods output ranked lists of windows with confidence scores. We try different ways to compute their PR curves, such as score thresholding and rank thresholding, with or without greedy NMS, and we report their best performance. For SC and LBI, rank thresholding without NMS (*i.e.* output all windows above a rank) gives consistently better AP scores. Note that SC and LBI already diversify their output windows, and their confidence scores are not calibrated across images. For MBox, applying score thresholding and NMS with the IOU threshold set at 0.4 provides the best performance.

We denote our full model as SalCNN+MAP. We also evaluate two baseline methods, SalCNN+NMS and SalCNN+MMR. SalCNN+NMS generates the detections by simply applying score thresholding and greedy NMS on our proposal windows. The IOU threshold for NMS is set at 0.4, which optimizes its AP scores. SalCNN+MMR uses the Maximum Marginal Relevance (MMR) measure to rescore the proposals [8, 3]. The new score of each proposal is computed as the original score minus a redundancy measure *w.r.t.* the previously selected proposals. We optimize the parameter for MMR and use score thresholding to compute the AP scores (see our supplementary material for more details). Moreover, we apply our optimization formulation (without tuning the parameters) and other baseline methods (with parameters optimized) on the raw outputs of MBox. In doing so, we can test how our MAP formulation generalizes to a different proposal model.

Evaluation Datasets. We evaluate our method mainly on three benchmark salient object datasets: MSO [53], DUT-O [51] and MSRA [29].

MSO contains many background images with no salient object and multiple salient objects. Each object is annotated separately. Images in this dataset are from the testing split of the SOS dataset [53].

DUT-O provides raw bounding box annotations of salient objects from five subjects. Images in this dataset can contain multiple objects, and a single annotated bounding box sometimes covers several nearby objects. We consolidate the annotations from five subjects to generate ground truth for evaluation (see supplementary material for details).

MSRA comprises 5000 images, each containing one dominant salient object. This dataset provides raw bounding boxes from nine subjects, and we consolidate these annotations in the same was as in DUT-O.

For completeness, we also report evaluation results on PASCAL VOC07 [18], which is originally for benchmarking object recognition methods. This dataset is not very suitable for our task, as it only annotates 20 categories of objects, many of which are not salient. However, it has been used for evaluating salient object detection in [21, 43]. As in [21, 43], we use all the annotated bounding boxes in VOC07 as class-agnostic annotations of salient objects.

4.1. Results

The PR curves of our method, baselines and other compared methods are shown in Fig. 3. The full AP scores are reported in Table 1. Our full model SalCNN+MAP significantly outperforms previous methods on MSO, DUT-O and MSRA. In particular, our method achieves about 15%, 34% and 20% relative improvement in AP over the best previous method MBox+NMS on MSO, DUT-O and MSRA respectively. This indicates that our model generalizes well to different datasets, even though it is only trained on the SOS training set. On VOC07, our method is slightly worse than MBox+NMS. Note that VOC07 is designed for object recognition, and MBox is optimized for this dataset [17]. We find that our method usually successfully detects the salient objects in this dataset, but often misses annotated objects in the background. Sample results are show in Fig. 5. More results can be found in our supplementary material.

Our MAP formulation consistently improves over the baseline methods NMS and MMR across all the datasets for both SalCNN and MBox. On average, our MAP attains more than 11% relative performance gain in AP over MMR for both SalCNN and MBox, and about about 12% (*resp.*

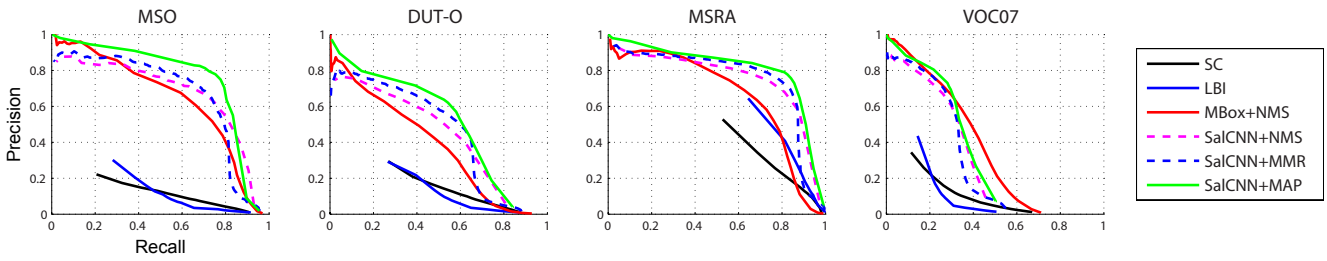

Figure 3: Precision-Recall curves. Our full method SalCNN+MAP significantly outperforms the other methods on MSO, DUT-O and MSRA. On VOC07, our method is slightly worse than MBox [46], but VOC07 is not a salient object dataset.

Table 1: AP scores. The best score on each dataset is shown in bold font, and the second best is underlined.

	MSO	DUT-O	MSRA	VOC07	Avg.
SC[21]	.121	.156	.388	.106	.194
LBI[43]	.144	.143	.513	.106	.226
MBox[46]+NMS	.628	.382	.647	.374	.508
MBox[46]+MMR	.595	.358	.578	.332	.466
MBox[46]+MAP	.644	.412	.676	**.394**	.532
SalCNN+NMS	.654	.432	<u>.722</u>	.300	.527
SalCNN+MMR	<u>.656</u>	<u>.447</u>	.716	.301	<u>.530</u>
SalCNN+MAP	**.734**	**.510**	**.778**	<u>.337</u>	**.590**

Table 2: AP scores in identifying background images on MSO.

SalCNN+MAP	SalCNN	MBox+MAP	MBox	LBI	SC
.89	.88	.74	.73	.27	.27

5%) relative performance gain over NMS for SalCNN (*resp.* MBox). Compared with NMS, the performance gain of our optimization method is more significant for SalCNN, because our adaptive region sampling method introduces extra proposal noise in the proposal set (see discussion in Section 3.4). The greedy NMS is quite sensitive to such noise, while our subset optimization formulation can more effectively handle it.

Detecting Background Images. Reporting the nonexistence of salient objects is an important task by itself [53, 49]. Thus, we further evaluate how our method and the competing methods handle background/cluttered images that do not contain any salient object. A background image is implicitly detected if there is no detection output by an algorithm. Table 2 reports the AP score of each method in detecting background images. The AP score of our full model SalCNN+MAP is computed by varying the parameter λ specified before. For SC, LBI, MBox and our proposal model SalCNN, we vary the score threshold to compute their AP scores.

As shown in Table 2, the proposal score generated by SC and LBI is a poor indicator of the existence of salient objects, since their scores are not calibrated across images. MBox significantly outperforms SC and LBI, while our proposal model SalCNN achieves even better performance,

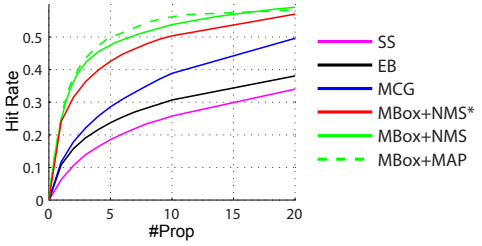

Figure 4: Object proposal generation performance (hit rate *vs.* average #Prop per image) on VOC07. Our MAP-based formulation further improves the state-of-the-art MBox method when #Prop is small.

which is expected as we explicitly trained our CNN model to suppress detections on background images. Our MAP formulation further improves the AP scores of SalCNN and MBox by 1 point.

Generating Compact Object Proposals. Object proposal generation aims to attain a high hit rate within a small proposal budget [24]. When a compact object proposal set is favored for an input image (*e.g.* in applications like weakly supervised localization [15, 44]), how proposals are filtered can greatly affect the hit rate. In Fig. 4, we show that using our subset optimization formulation can help improve the hit rate of MBox [46] when the average proposal number is less than 15 (see MBox+MAP vs MBox+NMS in Fig. 4). The performance of MBox using rank thresholding[2] (MBox+NMS*), together with SS [47], EB [56] and MCG [3], is also displayed for comparison.

4.2. Component Analysis

Now we conduct further analysis of our method on the MSO dataset, to evaluate the benefits of the main components of our system.

Adaptive Region Sampling. We compare our full model with two variants: the model without region sampling (w/o RS) and the model using uniform region sampling (Unif. RS) [17]. For uniform sampling, we extract

[2]Rank thresholding means outputting a fixed number of proposals for each image, which is a default setting for object proposal methods like SS, EB and MCG, as their proposal scores are less calibrated across images.

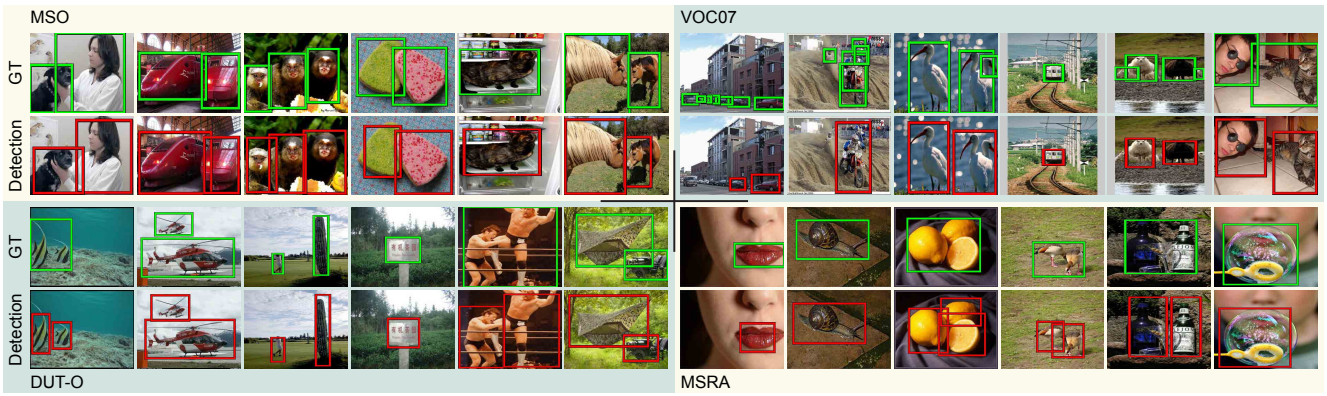

Figure 5: Sample detection results of our method when $\lambda = 0.1$. In the VOC07 dataset, many background objects are annotated, but our method only detects dominant objects in the scene. In the DUT-O and MSRA datasets, some ground truth windows cover multiple objects, while our method tends to localize each object separately. Note that we are showing all the detection windows produced by our method. More detection results are included in the supplementary material.

Table 3: AP scores of variants of our method. *Reg. Samp.* refers to variants with different region sampling strategies. *Win. Filtering* refers to variants using different window filtering methods. See text for details.

	Full Model	Reg. Samp.		Win. Filtering		
		w/o RS	Unif RS	Rank Thresh	Score Thresh	MMR
Overall	**.734**	.504	.594	.448	.654	.656
with Obj.	**.747**	.513	.602	.619	.668	.675
Single Obj.	**.818**	.676	.671	.717	.729	.721
Multi. Obj.	**.698**	.338	.540	.601	.609	.620
Large Obj.	**.859**	.790	.726	.776	.833	.804
Small Obj.	**.658**	.253	.498	.488	.558	.567

five sub-windows of 70% width and height of the image, by shifting the sub-window to the four image corners and the image center. The AP scores of our full model and these two variants are displayed in Table 3. Besides the AP scores computed over the whole MSO dataset, we also include the results on five subsets of images for more detailed analysis: 1) 886 images with salient objects, 2) 611 images with a single salient object, 3) 275 images with multiple salient objects, 4) 404 images with all small objects and 5) 482 images with a large object. An object is regarded as small if its bounding box occupies less than 25% area of the image. Otherwise, the object is regarded as large.

The best scores of the two variants are shown in red. The model with uniform region sampling generally outperforms the one without region sampling, especially on images with all small objects or multiple objects. However, on images with a large object, uniform region sampling worsens the performance, as it may introduce window proposals that are only locally salient, and it tends to cut the salient object. The proposed adaptive region sampling substantially enhances the performance on all the subsets of images, yielding over 20% relative improvement on the whole dataset.

MAP-based Subset Optimization. To further analyze our subset optimization formulation, we compare our full model with three variants that use different window filtering strategies. We evaluate the rank thresholding baseline (Rank Thresh in Table 3) and the score thresholding baseline (Score Thresh in Table 3) with the greedy NMS applied. We also evaluate the Maximum Marginal Relevance basline (MMR in Table 3) as in the previous experiment.

The results of this experiment are shown in Table 3. Our full model consistently gives better AP scores than all of the baselines, across all subsets of images. Even on constrained images with a single salient object, our subset optimization formulation still provides 12% relative improvement over the best baseline (shown in red in Table 3). This shows the robustness of our formulation in handling images with varying numbers of salient objects.

5. Conclusion

We presented a salient object detection system for unconstrained images, where each image may contain any number of salient objects or no salient object. A CNN model was trained to produce scored window proposals, and an adaptive region sampling method was proposed to enhance its performance. Given a set of scored proposals, we presented a MAP-based subset optimization formulation to jointly optimize the number and locations of detection windows. The proposed optimization formulation provided significant improvement over the baseline methods on various benchmark datasets. Our full method outperformed the state-of-the-art by a substantial margin on three challenging salient object datasets. Further experimental analysis validated the effectiveness of our system.

Acknowledgments. This research was supported in part by US NSF grants 0910908 and 1029430, and gifts from Adobe and NVIDIA.

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *CVPR*, 2009.

[2] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *PAMI*, 34(11):2189–2202, 2012.

[3] P. Arbelaez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In *CVPR*, 2014.

[4] J. Ba, V. Mnih, and K. Kavukcuoglu. Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*, 2014.

[5] O. Barinova, V. Lempitsky, and P. Kholi. On detection of multiple object instances using hough transforms. *PAMI*, 34(9):1773–1784, 2012.

[6] A. Borji, M.-M. Cheng, H. Jiang, and J. Li. Salient object detection: A benchmark. *ArXiv e-prints*, 2015.

[7] N. Buchbinder, M. Feldman, J. Naor, and R. Schwartz. A tight linear time (1/2)-approximation for unconstrained submodular maximization. In *Foundations of Computer Science*, 2012.

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

[9] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai. Fusing generic objectness and visual saliency for salient object detection. In *ICCV*, 2011.

[10] X. Chen and A. Gupta. Webly supervised learning of convolutional networks. In *ICCV*, 2015.

[11] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu. Global contrast based salient region detection. *PAMI*, 37(3):569–582, 2015.

[12] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. H. S. Torr. BING: Binarized normed gradients for objectness estimation at 300fps. In *CVPR*, 2014.

[13] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[14] C. Desai, D. Ramanan, and C. C. Fowlkes. Discriminative models for multi-class object layout. *IJCV*, 95(1):1–12, 2011.

[15] T. Deselaers, B. Alexe, and V. Ferrari. Weakly supervised localization and learning with generic knowledge. *IJCV*, 100(3):275–293, 2012.

[16] R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222, 1995.

[17] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. Scalable object detection using deep neural networks. In *CVPR*, 2014.

[18] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[19] U. Feige, V. S. Mirrokni, and J. Vondrak. Maximizing non-monotone submodular functions. *SIAM Journal on Computing*, 40(4):1133–1153, 2011.

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

[21] J. Feng, Y. Wei, L. Tao, C. Zhang, and J. Sun. Salient object detection by composition. In *ICCV*, 2011.

[22] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[23] V. Gopalakrishnan, Y. Hu, and D. Rajan. Random walks on graphs to model saliency in images. In *CVPR*, 2009.

[24] J. Hosang, R. Benenson, and B. Schiele. How good are detection proposals, really? In *BMVC*, 2014.

[25] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li. Salient object detection: A discriminative regional feature integration approach. In *CVPR*, 2013.

[26] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[27] H. Larochelle and G. E. Hinton. Learning to combine foveal glimpses with a third-order boltzmann machine. In *NIPS*, 2010.

[28] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille. The secrets of salient object segmentation. In *CVPR*, 2014.

[29] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. *PAMI*, 33(2):353–367, 2011.

[30] S. Lu, V. Mahadevan, and N. Vasconcelos. Learning optimal seeds for diffusion-based salient object detection. In *CVPR*, 2014.

[31] Y. Lu, W. Zhang, H. Lu, and X. Xue. Salient object detection using concavity context. In *ICCV*, 2011.

[32] Y. Luo, J. Yuan, P. Xue, and Q. Tian. Saliency density maximization for object detection and localization. In *ACCV*. 2011.

[33] R. Mairon and O. Ben-Shahar. In *A closer look at context: From coxels to the contextual emergence of object saliency*, 2014.

[34] L. Marchesotti, C. Cifarelli, and G. Csurka. A framework for visual saliency detection with applications to image thumbnailing. In *ICCV*, 2009.

[35] V. Mnih, N. Heess, A. Graves, et al. Recurrent models of visual attention. In *NIPS*, 2014.

[36] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[37] R. Rothe, M. Guillaumin, and L. Van Gool. Non-maximum suppression for object detection by passing messages between windows. In *ACCV*, 2014.

[38] S. Rujikietgumjorn and R. T. Collins. Optimized pedestrian detection for multiple and occluded people. In *CVPR*, 2013.

[39] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *arXiv preprint arXiv:1409.0575*, 2014.

[40] C. Scharfenberger, S. L. Waslander, J. S. Zelek, and D. A. Clausi. Existence detection of objects in images for robot vision using saliency histogram features. In *International Conference on Computer and Robot Vision*, 2013.

[41] X. Shen and Y. Wu. A unified approach to salient object detection via low rank matrix recovery. In *CVPR*, 2012.

[42] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[43] P. Siva, C. Russell, T. Xiang, and L. Agapito. Looking beyond the image: Unsupervised learning for object saliency and detection. In *CVPR*, 2013.

[44] P. Siva and T. Xiang. Weakly supervised object detector learning with model drift detection. In *ICCV*, pages 343–350, 2011.

[45] B. Suh, H. Ling, B. B. Bederson, and D. W. Jacobs. Automatic thumbnail cropping and its effectiveness. In *ACM symposium on User interface software and technology*, 2003.

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

[47] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *IJCV*, 104(2):154–171, 2013.

[48] R. Valenti, N. Sebe, and T. Gevers. Image saliency by isocentric curvedness and color. In *ICCV*, 2009.

[49] P. Wang, J. Wang, G. Zeng, J. Feng, H. Zha, and S. Li. Salient object detection for searched web images via global saliency. In *CVPR*, 2012.

[50] Q. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical saliency detection. In *CVPR*, 2013.

[51] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013.

[52] G. Yildirim and S. Ssstrunk. FASA: Fast, Accurate, and Size-Aware Salient Object Detection. In *ACCV*, 2014.

[53] J. Zhang, S. Ma, M. Sameki, S. Sclaroff, M. Betke, Z. Lin, X. Shen, B. Price, and R. Měch. Salient object subitizing. In *CVPR*, 2015.

[54] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech. Minimum barrier salient object detection at 80 fps. In *ICCV*, 2015.

[55] J.-Y. Zhu, J. Wu, Y. Wei, E. Chang, and Z. Tu. Unsupervised object class discovery via saliency-guided multiple class learning. In *CVPR*, 2012.

[56] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014.