The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Weakly Supervised Salient Object Detection Using Image Labels

Guanbin Li,[1] **Yuan Xie,**[1] **Liang Lin**[1,2*]
[1]School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China
[2]SenseTime Group Limited

Abstract

Deep learning based salient object detection has recently achieved great success with its performance greatly outperforms any other unsupervised methods. However, annotating per-pixel saliency masks is a tedious and inefficient procedure. In this paper, we note that superior salient object detection can be obtained by iteratively mining and correcting the labeling ambiguity on saliency maps from traditional unsupervised methods. We propose to use the combination of a coarse salient object activation map from the classification network and saliency maps generated from unsupervised methods as pixel-level annotation, and develop a simple yet very effective algorithm to train fully convolutional networks for salient object detection supervised by these noisy annotations. Our algorithm is based on alternately exploiting a graphical model and training a fully convolutional network for model updating. The graphical model corrects the internal labeling ambiguity through spatial consistency and structure preserving while the fully convolutional network helps to correct the cross-image semantic ambiguity and simultaneously update the coarse activation map for next iteration. Experimental results demonstrate that our proposed method greatly outperforms all state-of-the-art unsupervised saliency detection methods and can be comparable to the current best strongly-supervised methods training with thousands of pixel-level saliency map annotations on all public benchmarks.

Introduction

Salient object detection is designed to accurately detect distinctive regions in an image that attract human attention. Recently, this topic has attracted widespread interest in the research community of computer vision and cognitive science as it can be applied to benefit a wide range of artificial intelligence and vision applications, such as robot intelligent control (Shon et al. 2005), content-aware image editing (Avidan and Shamir 2007), visual tracking (Mahadevan and Vasconcelos 2009) and video summarization (Ma et al. 2002).

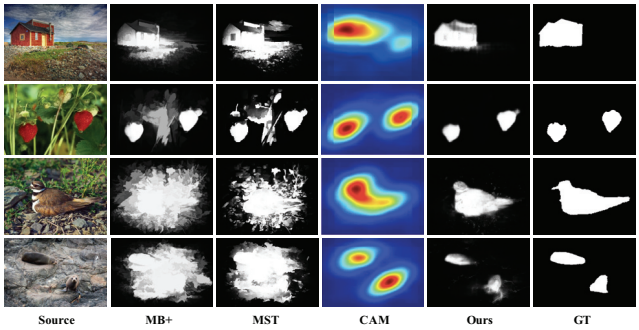

Figure 1: Two kinds of defects in state-of-the-art unsupervised salient object detection methods vs. the results of our proposed weakly supervised optimization framework.

Recently, the deployment of deep convolutional neural networks has resulted in significant progress in salient object detection (Li and Yu 2016b; 2016a; Liu and Han 2016; Wang et al. 2016). The performance of these CNN based methods, however, comes at the cost of requiring pixel-wise annotations to generate training data. For salient object detection, it is painstaking to annotate mask-level label and takes several minutes for an experienced annotator to label one image. Moreover, as the definition of an object being salient is very subjective, there often exists multiple diverse annotations for a same image between different annotators. To ensure the quality of training data sets, these images with ambiguous annotations should be removed, which makes the labeling task more laborious and time-consuming. This time-consuming task is bound to limit the total amount of pixel-wise training samples and thus become the bottleneck of further development of fully-supervised learning based methods.

As a low level vision problem, there exists an ocean of unsupervised salient object detection methods (Wei et al. 2012; Cheng et al. 2015; Tu et al. 2016; Zhang et al. 2015; Yang et al. 2013). These methods are usually based on low-level features such as color, gradient or contrast and some saliency priors, such as the center prior (Liu et al. 2011) and the background prior (Wei et al. 2012). As it is impracti-

*Corresponding author is Liang Lin (Email: linliang@ieee.org).
This work was supported in part by the National Natural Science Foundation of China under Grant 61702565, in part by the Special Program of the NSFC-Guangdong Joint Fund for Applied Research on Super Computation (the second phase), in part by Guangdong Natural Science Foundation Project for Research Teams under Grant 2017A030312006. This work was also sponsored by CCF-Tencent Open Research Fund.

cal to define a set of universal rules for how an object being salient, each of these bottom-up methods works well for some images, but none of them can handle all the images. By observing the failure cases, We found that most of the saliency detection error lies in the lack of spatial correlation inference and image semantic contrast detection. As shown in Figure 1, the two unsupervised saliency detection methods are only able to detect part of the salient objects in the first two cases as they fail to take into account the spatial consistency (e.g. encouraging nearby pixels with similar colors to take similar saliency scores), while for the last two cases, the two methods completely fail to detect the salient objects as these objects are of very low contrast in terms of low-level features (they are salient in high semantic contrast). These two kinds of failure cases are hard to be found in fully-supervised methods.

During the training and testing of several fully-supervised saliency detection methods (Li and Yu 2016a; Liu and Han 2016; Wang et al. 2016), we found that a well-trained deep-convolution network without over-fitting can even correct some user annotation error exists in the training samples. We conjecture that a large amount of model parameters contained in deep neural network can be trained to discover the universal rules implied in large scale training samples and thus can help to detect the ambiguity in the annotation mask (noisy annotation) and figure out a "correct" one which being in line with the hidden rules. Moreover, recent work has shown that CNNs being trained on image-level labels for classification have remarkable ability to localize the most discriminative region of an image (Zhou et al. 2016).

Inspired by these observations, in this paper, we address the weakly supervised salient object detection task using only image-level labels, which in most cases specify salient objects within the image. We develop an algorithm that exploits saliency maps generated from any unsupervised method as noisy annotations to train convolutional networks for better saliency maps. Specifically, we first propose a conditional random field based graphical model to correct the internal label ambiguity by enhancing the spatial coherence and salient object localization. Meanwhile, a multi-task fully convolutional ResNet (He et al. 2015) is learned, which is supervised by the iteratively corrected pixel-level annotations as well as image labels (indicating significant object class within an image), and in turn provides two probability maps to generate an updated unary potential for the graphical model. The first probability map is called Class Activation Map (CAM) and it highlights the discriminative object parts detected by the image classification-trained CNN while the second one being a more accurate saliency map trained from pixel-wise annotation. Though CAM itself is a relatively coarse pixel-level probability map, it shows very accurate salient object localization ability and thus can be used as a guide to generate more precise pixel-wise annotation for a second round training. The proposed method is optimized alternately until a stopping criteria appears. In our experiment, we find that although CAM is trained using images from a fix number of image classes, it generalizes well to images of unknown categories, resulting in an intensely accurate salient object positioning for generic salient objects. The proposed optimization framework also theoretically applies to all unsupervised salient object detection methods and is able to generate more accurate saliency map very efficiently in fewer than one second per image no matter how time-consuming the original model.

In summary, this paper has the following contributions:

- We introduce a generic alternate optimization framework to fill the performance gap between supervised and unsupervised salient object detection methods without resorting to laborious pixel labeling.
- We propose a conditional random field based graphical model to cleanse the noisy pixel-wise annotation by enhancing the spatial coherence as well as salient object localization.
- We also design a multi-task fully convolutional ResNet-101 to both generate a coarse class activation map (CAM) and a pixel-wise saliency probability map, the cooperation of which can help to detect and correct the cross-image annotation ambiguity, generating more accurate saliency annotation for iterative training.

Alternate Saliency Map Optimization

As shown in Figure 2, our proposed saliency map optimization framework consists of two components, a multi-task fully convolutional network (Multi-FCN) and a graphical model based on conditional random fields (CRF). Given the Microsoft COCO dataset (Lin et al. 2014) with multiple image labels corresponding to each image, we initially utilize a state-of-the-art unsupervised salient object detection method, i.e. minimum barrier salient object detection (MB+), to generate the saliency maps of all training images. The produced saliency maps as well as their corresponding image labels are employed to train the Multi-FCN, which simultaneously learns to predict a pixel-wise saliency map and an image class distribution. When training converged, a class activation mapping technique (Zhou et al. 2016) is applied to the Multi-FCN to generate a serious of class activation maps (CAMs). Then the initial saliency map, the predicted saliency map from Multi-FCN as well as the average map of the top three CAMs (CAM prediction corresponding to top 3 classes) are employed to the CRF model to get the corresponding maps with better spatial coherence and contour localization. We further propose an annotation updating scheme to construct new saliency map annotations from these three maps with CRF for a second iteration of Multi-FCN training. Finally, to generalize the model for saliency detection of unknown image labels, we further finetune the saliency map prediction stream of the Multi-FCN guided by generated CAM using salient object detection datasets (e.g. MSRA-B and HKUIS) without annotations.

Multi-Task Fully Convolutional Network

In the multi-task fully convolutional stream, we aim to design an end-to-end convolutional network that can be viewed as a combination of the image classification task and the pixel-wise saliency prediction task. To conceive such an

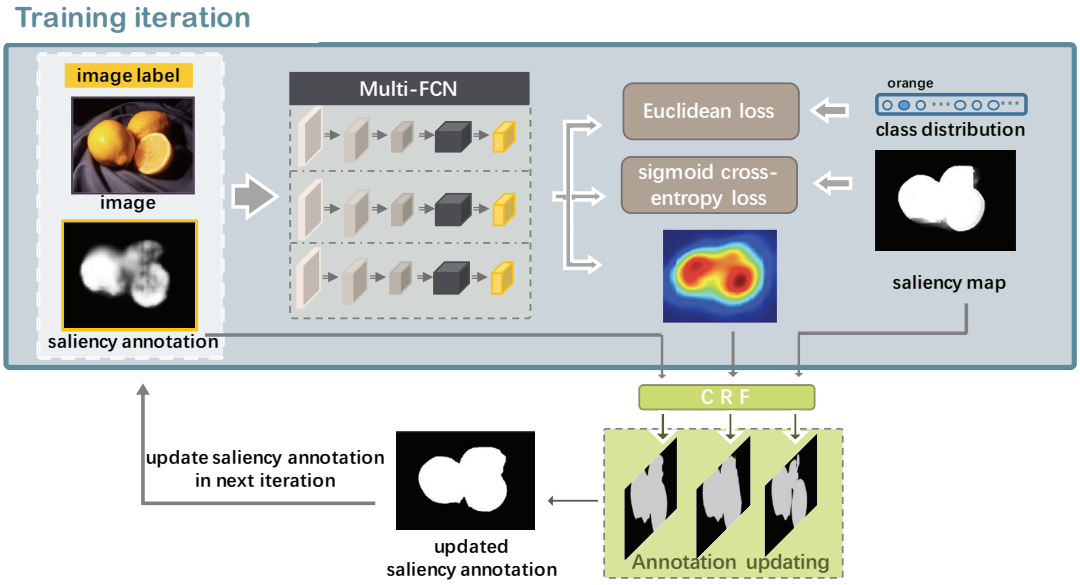

Figure 2: Overall framework for alternate saliency map optimization.

end-to-end architecture, we have the following considerations. First, the network should be able to correct the noisy initial saliency annotations as well as possible by mining the semantic ambiguity between the images. Second, the network should be able to be end to end trainable to output a saliency map with appropriate resolution. Last but not the least, it should also be able to detect visual contrast at different scales.

We choose ResNet-101 (He et al. 2015) as our pre-trained network and modify it to meet our requirements. We first refer to (Chen et al. 2014) and re-purpose it into a dense image saliency prediction network by replacing its 1000-way linear classification layer with a linear convolutional layer with a 1×1 kernel and two output channels. The feature maps after the final convolutional layer is only $1/32$ of that of the original input image because the original ResNet-101 consists of one pooling layer and 4 convolutional layers, each of which has stride 2. We call these five layers down-sampling layers. As described in (He et al. 2015), the 101 layers in ResNet-101 can be divided into five groups. Feature maps computed by different layers in each group share the same resolution. To make the final saliency map denser, we skip subsampling in the last two down-sampling layers by setting their stride to 1, and increase the dilation rate of subsequent convolutional kernels using the dilation algorithm to enlarge their receptive fields as (Chen et al. 2014). Therefore, all the features maps in the last three groups have the same resolution, $1/8$ original resolution after network transformation.

As it has been widely verified that feeding multiple scales of an input image to networks with shared parameters are rewarding for accurately localizing objects of different scales (Chen et al. 2015; Lin et al. 2015), we replicate the fully convolutional ResNet-101 network three times, each responsible for one input scale s ($s \in \{0.5, 0.75, 1\}$). Each scale s of the input image is fed to one of the three replicated ResNet-101, and outputs a two-channel probability map in the resolution of scale s, denoted as M_c^s, where $c \in \{0, 1\}$ denotes the two classes for saliency detection. The three probability maps are resized to the same resolution as the raw input image using bilinear interpolation, summed up and fed to a sigmoid layer to produce the final probability map. The network framework is shown in Figure 3.

For image classification task, as we desire to perform object localization from the classification model, we refer to (Zhou et al. 2016) and integrate a global average pooling layer for generating class activation maps. Specifically, as shown in Figure 3, we rescale the three output feature maps of the last original convolutional layer in ResNet-101 (corresponds to three input scale) to the same size ($1/8$ original resolution) and concatenate to form feature maps for classification. We further perform global average pooling on the concatenated convolutional feature maps and use those as features for a fully-connected layer which produces the desired classes distribution output. Let $f_k(x, y)$ represent the activation of channel k in the concatenated feature map at spatial location (x, y). Define M_c as the class activation map for class c, where each spatial element can be calculated as follows (Zhou et al. 2016):

$$M_c(x, y) = \sum_k w_k^c f_k(x, y). \quad (1)$$

w_k^c is the weight corresponding to class c for unit k (after global average pooling, each channel of the concatenated feature map becomes a unit activation value).

Graphical Model for Saliency Map Refinement

By observing the saliency maps generated by state-of-the-art unsupervised methods, we find that for images with low contrast and complex background, the salient object can hardly be completely detected, with common defects exist

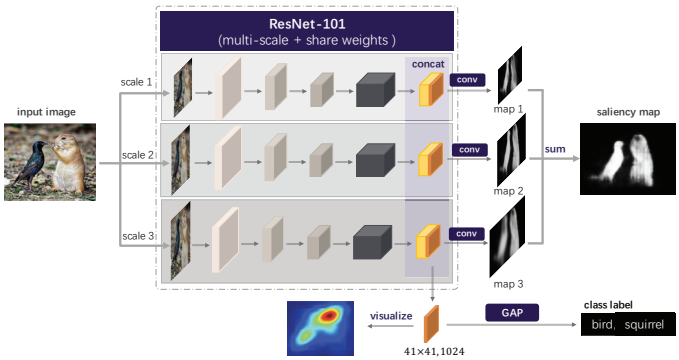

Figure 3: The architecture of our multi-task fully convolutional network (Multi-FCN).

in spatial consistency and contour preserving. We call these defects internal labeling ambiguity in noisy saliency annotations. Fully connect CRF model has been widely used in semantic segmentation (Krähenbühl and Koltun 2012; Chen et al. 2014) to both refine the segmentation result and better capture the object boundaries. It has also been used as a post-processing step in (Li and Yu 2016a; 2016b) for saliency map refinement. In this paper, we refer to (Li and Yu 2016a) and utilize the same formulation and solver of the two classes fully connected CRF model to correct the internal labeling ambiguity. The output of the CRF operation is a probability map, the value of which denotes the probability of each pixel being salient. We convert it into a binary label by thresholding when being used as training sample annotations.

Saliency Annotations Updating Scheme

We denote the original input image as I and the corresponding saliency map of the specific unsupervised method as S_{anno}. After convergence of the first complete training of Multi-FCN, we apply the trained model to generate saliency maps as well as the average map of the top 3 class activation maps for all training images. We denote the predicted saliency map as $S_{predict}$ and the average class activation map as S_{cam}. Furthermore, we also perform fully connected CRF operation to the initial saliency maps produced by a specific unsupervised method, the predicted saliency map from Multi-FCN as well as the average class activation map. The resulting saliency maps are denoted as C_{anno}, $\mathrm{C}_{predict}$ and C_{cam} respectively. Base on this, we update the training samples as well as their corresponding saliency annotations for the next iteration according to Algorithm 1. $\mathrm{CRF}\left(\right)$ denotes the CRF operation while S_{update} refers to the updated saliency map annotation, which is further used as the saliency groundtruth for the next iterative training. $\mathrm{MAE}\left(\right)$ is defined as the average pixelwise absolute difference between two saliency maps (i.e. S_1 and S_2), which is calculated as follows:

$$\mathrm{MAE}\left(S_1, S_2\right) = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |S_1(x,y) - S_2(x,y)|. \tag{2}$$

where W and H being the width and height of the saliency map.

Algorithm 1 Saliency Annotations Updating

Require: Current saliency map annotation S_{anno}, the predicted saliency map $S_{predict}$, CRF output of current saliency map annotation C_{anno}, CRF output of the predicted saliency map $\mathrm{C}_{predict}$ and CRF output of the class activation map C_{cam}

Ensure: The updated saliency map annotation S_{update}.

if $\mathrm{MAE}\left(\mathrm{C}_{anno}, \mathrm{C}_{predict}\right) \leq \alpha$ **then**
  $S_{update} = \mathrm{CRF}\left(\frac{S_{anno}+S_{predict}}{2}\right)$
else if $\mathrm{MAE}\left(\mathrm{C}_{anno}, \mathrm{C}_{cam}\right) > \beta$ and $\mathrm{MAE}\left(\mathrm{C}_{predict}, \mathrm{C}_{cam}\right) > \beta$ **then**
  Discard the training sample in next iteration
else if $\mathrm{MAE}\left(\mathrm{C}_{anno}, \mathrm{C}_{cam}\right) \leq \mathrm{MAE}\left(\mathrm{C}_{predict}, \mathrm{C}_{cam}\right)$ **then**
  $S_{update} = \mathrm{C}_{anno}$
else
  $S_{update} = \mathrm{C}_{predict}$
end if

Multi-FCN Training with Weak Labels

The training of the weakly supervised saliency map optimization framework is composed of two stages, both of which are based on an alternative training scheme. In the first stage, we train the model using Microsoft COCO dataset (Lin et al. 2014) with multiple image labels per image. Firstly, we choose a state-of-the-art unsupervised salient object detection model and apply it to produce an initial saliency map for each image in the training. Then we simply put the saliency maps as initial annotation and train the Multi-FCN for a pixel-wise saliency prediction as well as the classification model for better class activation map generation. While training, we validate on the validation set of the COCO dataset also with generated pixel-wise noisy annotations being the groundtruth. Also note that in order to speed up the training, we initialize the Multi-FCN with a pre-trained model over the ImageNet dataset (Deng et al. 2009) instead of training from scratch. After training convergence, we choose the model with lowest validation error as the final model for this iteration, and apply it to generate saliency maps as well as the average map of top 3 class activation maps for all training images. Secondly, we apply saliency annotations updating scheme according to Section to create updated training tuples (images, saliency annotation and image label) for a second round of training. We alternately train the model until a stopping criteria appears. After each training round, we evaluate the mean MAE between each pair of saliency annotation (Pseudo Groundtruth) and the predicted saliency map, and the stopping criteria is defined to be the mean MAE gets lower than a specific threshold or the total number of training rounds reaches 5. (Noted that as being a weakly supervised method, we do not use true annotations to determine the merits of the model).

Finally, in order to generalize the model for generic salient object detection with unknown image labels, we further finetune the saliency map prediction stream of the Multi-FCN guided by offline CAMs using salient object detection datasets (e.g. the training images of MSRA-B and HKU-IS) without annotations, until the stopping criteria appears. Here, we calculate the mean of the top 5 CAMs as the guided CAM, and we discover that although CAM is trained with specific image classification labels, its predicted CAMs of the most similar categories in the category set can still highlight the most discriminative regions in the image and thus still works well as an auxiliary guidance for generic salient object detection. The loss function for updating Multi-FCN for pixel-wise saliency prediction is defined as the sigmoid cross entropy between the generated ground truth (G) and the predicted saliency map (S):

$$\begin{aligned} L = &- \beta_i \sum_{i=1}^{|I|} G_i \log P\left(S_i = 1 | I_i, W\right) \\ &- \left(1 - \beta_i\right) \sum_{i=1}^{|I|} \left(1 - G_i\right) \log P\left(S_i = 0 | I_i, W\right), \end{aligned} \tag{3}$$

where W denotes the collection of corresponding network parameters in the Multi-FCN, β_i is a weight balancing the number of salient pixels and unsalient ones, and $|I|$, $|I|_-$ and $|I|_+$ denote the total number of pixels, unsalient pixels and salient pixels in image I, respectively. Then $\beta_i = \frac{|I|_-}{|I|}$ and $1 - \beta_i = \frac{|I|_+}{|I|}$. When training for multi-label object classification, we simply employ the Euclidean loss as the objective function and only update the parameters of the fully connected inference layer with parameters of the main ResNet-101 being unchanged.

Experimental Results

Implementation

Our proposed Multi-FCN has been implemented on the public DeepLab code base (Chen et al. 2014). A GTX Titan X GPU is used for both training and testing. As described in Section , the Multi-FCN involves two stages training. In the first stage, we train on Microsoft COCO object detection dataset for multi-label recognition, which comprises a training set of 82,783 images, and a validation set of 40,504 images. The dataset covers 80 common object categories, with about 3.5 object labels per image. In the second stage, we combine the training images of both the MSRA-B dataset (2500 images) (Liu et al. 2011) and the HKU-IS dataset (2500 images) (Li and Yu 2016b) as our training set (5000 images), with all original saliency annotations removed. The validation sets without annotations in the aforementioned two datasets are also combined as our validation set (1000 images). During training, the mini-batch size is set to 2 and we choose to update the loss every 5 iterations. We set the momentum parameter to 0.9 and the weight decay to 0.0005 for both subtasks. The total number of iteration is set to 8K during each training round. During saliency annotation updating, the thresholds α and β are set to 15 and 40 respectively. The mean MAE of the training stop criteria is set to 0.05 in our experiment.

Datasets

We conducted evaluations on six public salient object benchmark datasets: MSRA-B (Liu et al. 2011), PASCAL-S (Li et al. 2014), DUT-OMRON(Yang et al. 2013), HKU-IS (Li and Yu 2016b), ECSSD (Yan et al. 2013) and SOD (Martin et al. 2001). Though we do not use any user annotations in training, we get to know the training and validation sets of the MSRA-B and HKU-IS datasets in advance. Therefore, for the sake of fairness, we evaluate our model on the testing sets of these two datasets and on the combined training and testing sets of other datasets.

Evaluation Metrics

We adopt precision-recall curves (PR), maximum F-measure and mean absolute error (MAE) as our performance measures. The continuous saliency map is binarized using different thresholds varying from 0 to 1. At each threshold value, a pair of precision and recall value can be obtained by comparing the binarized saliency map against the binary groundtruth. The PR curve of a dataset is obtained from all pairs of average precision and recall over saliency maps of all images in the dataset. The F-measure is defined as $F_\beta = \frac{(1+\beta^2) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}$, where β^2 is set to 0.3. We report the maximum F-measure computed from all precision-recall pairs. MAE is defined as the average pixelwise absolute difference between the binary ground truth and the saliency map (Perazzi et al. 2012) as described in Equation 2.

Comparison with the Unsupervised State-of-the-Art

Our proposed alternate saliency map optimization framework requires an unsupervised benchmark model as initialization. In this section, we choose the state-of-the-art minimum barrier salient object detection (MB+) method as a baseline and take the optimized model as our final model when compared with other benchmarks. In Section , we will list more results of our proposed method on other baseline models to demonstrate the effectiveness of our proposed algorithm.

We compare our method with eight classic or state-of-the-art unsupervised saliency detection algorithms, including GS (Wei et al. 2012), SF (Perazzi et al. 2012), HS (Yan et al. 2013), MR (Yang et al. 2013), GC (Cheng et al. 2015), BSCA (Qin et al. 2015), MB+ (Zhang et al. 2015) and MST (Tu et al. 2016). For fair comparison, the saliency maps of different methods are provided by authors or obtained from the available implementations.

A visual comparison is given in Fig. 5. As can be seen, our method generates more accurate saliency maps in various challenging cases, e.g., object in complex background and low contrast between object and background. It is particularly noteworthy that our proposed method employed the saliency maps generated by MB+ (Zhang et al. 2015) as initial noisy annotations for iterative training, it can learn to

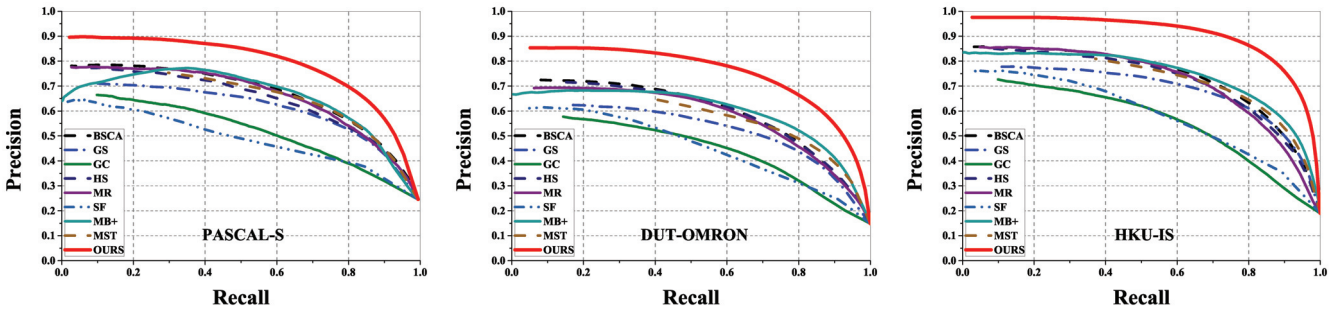

Figure 4: Comparison of precision-recall curves among 9 salient region detection methods on 3 datasets. Our proposed ASMO consistently outperforms other methods across all the testing datasets.

Data Set	Metric	GS	SF	HS	MR	GC	BSCA	MB+	MST	ASMO	ASMO+
MSRA-B	maxF	0.777	0.700	0.813	**0.824**	0.719	0.830	0.822	0.809	**0.890**	**0.896**
	MAE	0.144	0.166	0.161	0.127	0.159	0.130	0.133	**0.098**	**0.067**	**0.068**
ECSSD	maxF	0.661	0.548	0.727	0.736	0.597	**0.758**	0.736	0.724	**0.837**	**0.845**
	MAE	0.206	0.219	0.228	0.189	0.233	0.183	0.193	**0.155**	**0.110**	**0.112**
HKU-IS	maxF	0.682	0.590	0.710	0.714	0.588	0.723	**0.727**	0.707	**0.846**	**0.855**
	MAE	0.166	0.173	0.213	0.174	0.211	0.174	0.180	**0.139**	**0.086**	**0.088**
DUT-OMRON	maxF	0.556	0.495	0.616	0.610	0.495	0.617	**0.621**	0.588	**0.722**	**0.732**
	MAE	0.173	**0.147**	0.227	0.187	0.218	0.191	0.193	0.161	**0.101**	**0.100**
PASCAL-S	maxF	0.620	0.493	0.641	0.661	0.539	0.666	**0.673**	0.657	**0.752**	**0.758**
	MAE	0.223	0.240	0.264	0.223	0.266	0.224	0.228	**0.194**	**0.152**	**0.154**
SOD	maxF	0.620	0.516	0.646	0.636	0.526	0.654	**0.658**	0.647	**0.751**	**0.758**
	MAE	0.251	0.267	0.283	0.259	0.284	0.251	0.255	**0.223**	**0.185**	**0.187**

Table 1: Comparison of quantitative results including maximum F-measure (larger is better) and MAE (smaller is better). The best three results on each dataset are shown in **red**, **blue**, and **green** , respectively.

mine the ambiguity inside the original noisy labels and the semantic annotation ambiguity across different images, correct them, and eventually produce an optimized results far better than the original ones. As a part of quantitative evaluation, we show a comparison of PR curves in Fig. 4, as shown in the figure, our method significantly outperforms all state-of-the-art unsupervised salient object detection algorithms. Moreover, a quantitative comparison of maximum F-measure and MAE is listed in Table. 1. Our proposed alternate saliency map optimization (ASMO) improves the maximum F-measure achieved by the best-performing existing algorithm by 8.74%, 11.48%, 17.61%, 17.87%, 12.63% and 15.20% respectively on MSRA-B, ECSSD, HKU-IS, DUT-OMRON, PASCAL-S and SOD. And at the same time, it lowers the MAE by 31.63%, 29.03%, 38.13%, 31.97%, 21.65% and 17.04% respectively on MSRA-B, ECSSD, HKU-IS, DUT-OMRON, PASCAL-S and SOD. We also evaluate the performance of further applying dense CRF to our proposed method, listed as ASMO+ in the table.

Ablation Studies

Effectiveness of Alternate Saliency Map Optimization Our proposed Multi-FCN based saliency map optimization framework is composed of two components, a multi-task fully convolutional network (Multi-FCN) and a graphical model based on conditional random fields (CRF). To show the effectiveness of the proposed optimization method, we compare the saliency map S_1 generated from the original method, the saliency map S_2 from directly employing dense CRF to the original method, the saliency map S_3 from training Multi-FCN with generated saliency maps but without CRF or CAM guided, the saliency map S_4 from training Multi-FCN with generated saliency maps and CRF guided but without CAM and the saliency map S_5 from our full pipeline using DUT-OMRON dataset. As shown in Tab. 3, employing dense CRF operation and iterative training on fully convolutional ResNet-101 with original saliency map as noisy groundtruth can both boost the performance of the original unsupervised method. Our alternately updating scheme can integrate both of these two complementary advantages which further gains 5.22% improvement on maximum F-measure and 16.6% decrease on MAE. Moreover, CAM guided saliency annotations updating scheme plays a paramount role in our optimization framework which also greatly improve the saliency map performance.

Sensitivities to Benchmark Method Selection As described in Section , our proposed alternate saliency map optimization method is based on an unsupervised benchmark model as initialization. To demonstrate that our proposed method is widely applicable to the optimization of any unsupervised salient object detection method, we apply our optimization method to the other two recently published unsupervised saliency detection methods, including BSCA (Qin

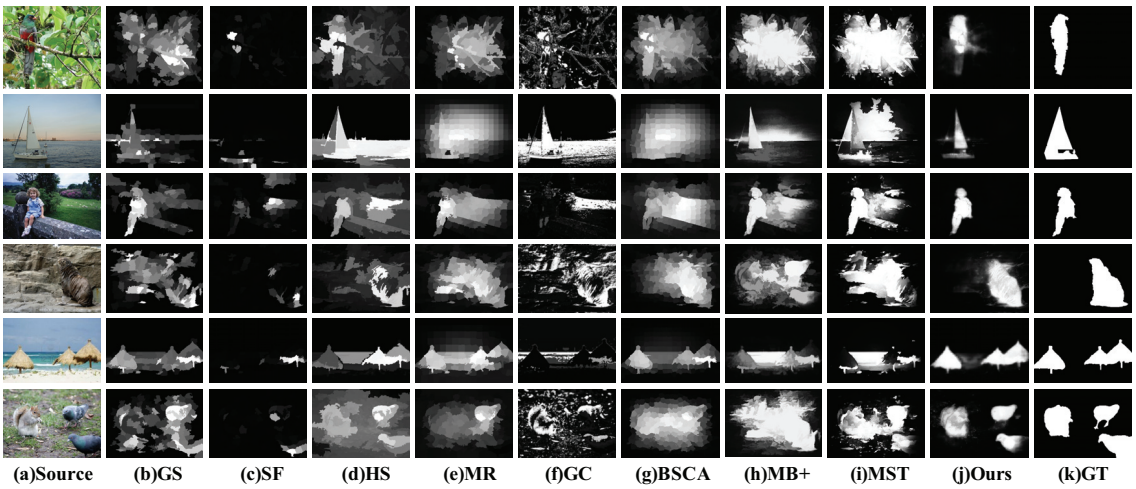

Figure 5: Visual comparison of saliency maps from state-of-the-art methods. The ground truth (GT) is shown in the last column. Our proposed method consistently produces saliency maps closest to the ground truth.

Table 2: Comparison with fully supervised salient object detection methods on HKU-IS, ECSSD and PASCAL-S datasets.

Data Set	Metric	DRFI	LEGS	MC	MDF	RFCN	DHSNet	DCL+	ASMO+	ASMO+ (with mask)
HKU-IS	maxF	0.776	0.770	0.798	0.861	0.896	0.892	0.904	0.855	**0.913**
	MAE	0.167	0.118	0.102	0.076	0.073	0.052	0.049	0.088	**0.041**
ECSSD	maxF	0.782	0.827	0.837	0.847	0.899	0.907	0.901	0.845	**0.918**
	MAE	0.170	0.118	0.100	0.106	0.091	0.059	0.068	0.112	**0.057**
PASCAL-S	maxF	0.690	0.752	0.740	0.764	0.832	0.824	0.822	0.758	**0.847**
	MAE	0.210	0.157	0.145	0.145	0.118	0.094	0.108	0.154	**0.092**

Table 3: Effectiveness evaluation of different components of alternate saliency map optimization on DUT-OMRON dataset.

Metric	S_1	S_2	S_3	S_4	S_5	S_5+CRF
maxF	0.588	0.630	0.651	0.685	0.722	0.732
MAE	0.161	0.178	0.151	0.126	0.101	0.100

Table 4: Evaluation of different benchmark methods in alternate saliency map optimization on DUT-OMRON dataset.

Metric	MB+	ASMO (MB+)	BSCA	ASMO (BSCA)	MST	ASMO (MST)
maxF	0.621	0.722	0.617	0.685	0.588	0.691
MAE	0.193	0.101	0.191	0.121	0.161	0.126

et al. 2015) and MST (Tu et al. 2016). Experimental results in Tab. 4 have shown that although these methods have achieved very good results, the application of our proposed optimization algorithm can still significantly improve their performance.

Evaluation on Semi-Supervised Setting In this section, we aim to compare our proposed method with the state-of-the-art fully supervised methods. As shown in Tab. 2, our proposed weakly supervised ASMO with CRF already consistently outperforms 3 fully supervised methods including DRFI (Jiang et al. 2013), LEGS (Wang et al. 2015) and MC (Zhao et al. 2015), and it is comparable to MDF (Li and Yu 2016b). Particularly noteworthy that when we add the groundtruth mask of the training set of MSRA-B dataset to form a semi-supervised setting of our method, it greatly outperforms all state-of-the-art fully supervised methods across all the three testing datasets (HKU-IS, ECSSD and PASCAL-S). We conjecture that our model considered more semantic information than existing fully-supervised models as we additionally included the Microsoft COCO dataset in our initial training.

Conclusions

In this paper, we have introduced a generic alternate optimization framework to improve the saliency map quality of any unsupervised salient object detection methods by alternately exploiting a graphical model and training a multi-task fully convolutional network for model updating. Experimental results demonstrate that our proposed method greatly outperforms all state-of-the-art unsupervised saliency detection methods and can be comparable to the current best strongly-supervised methods training with thousands of pixel-level saliency map annotations on all public benchmarks.

References

Avidan, S., and Shamir, A. 2007. Seam carving for content-aware image resizing. *ACM Transactions on graph-*

ics 26(3):10.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*.

Chen, L.-C.; Yang, Y.; Wang, J.; Xu, W.; and Yuille, A. L. 2015. Attention to scale: Scale-aware semantic image segmentation. *arXiv preprint arXiv:1511.03339*.

Cheng, M.-M.; Mitra, N. J.; Huang, X.; Torr, P. H.; and Hu, S.-M. 2015. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37(3):569–582.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, 248–255.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*.

Jiang, H.; Wang, J.; Yuan, Z.; Wu, Y.; Zheng, N.; and Li, S. 2013. Salient object detection: A discriminative regional feature integration approach. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2083–2090.

Krähenbühl, P., and Koltun, V. 2012. Efficient inference in fully connected crfs with gaussian edge potentials. *arXiv preprint arXiv:1210.5644*.

Li, G., and Yu, Y. 2016a. Deep contrast learning for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 478–487.

Li, G., and Yu, Y. 2016b. Visual saliency detection based on multiscale deep cnn features. *IEEE Transactions on Image Processing* 25(11):5012–5024.

Li, Y.; Hou, X.; Koch, C.; Rehg, J. M.; and Yuille, A. L. 2014. The secrets of salient object segmentation. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, 280–287.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, 740–755.

Lin, G.; Shen, C.; Reid, I.; et al. 2015. Efficient piecewise training of deep structured models for semantic segmentation. *arXiv preprint arXiv:1504.01013*.

Liu, N., and Han, J. 2016. Dhsnet: Deep hierarchical saliency network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 678–686.

Liu, T.; Yuan, Z.; Sun, J.; Wang, J.; Zheng, N.; Tang, X.; and Shum, H.-Y. 2011. Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(2):353–367.

Ma, Y.-F.; Lu, L.; Zhang, H.-J.; and Li, M. 2002. A user attention model for video summarization. In *Proceedings of the tenth ACM international conference on Multimedia*, 533–542.

Mahadevan, V., and Vasconcelos, N. 2009. Saliency-based discriminant tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1007–1013.

Martin, D.; Fowlkes, C.; Tal, D.; and Malik, J. 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of International Conference on Computer Vision*, volume 2, 416–423.

Perazzi, F.; Krähenbühl, P.; Pritch, Y.; and Hornung, A. 2012. Saliency filters: Contrast based filtering for salient region detection. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, 733–740.

Qin, Y.; Lu, H.; Xu, Y.; and Wang, H. 2015. Saliency detection via cellular automata. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, 110–119.

Shon, A. P.; Grimes, D. B.; Baker, C. L.; Hoffman, M. W.; Zhou, S.; and Rao, R. P. 2005. Probabilistic gaze imitation and saliency learning in a robotic head. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2865–2870.

Tu, W.-C.; He, S.; Yang, Q.; and Chien, S.-Y. 2016. Real-time salient object detection with a minimum spanning tree. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2334–2342.

Wang, L.; Lu, H.; Ruan, X.; and Yang, M.-H. 2015. Deep networks for saliency detection via local estimation and global search. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, 3183–3192.

Wang, L.; Wang, L.; Lu, H.; Zhang, P.; and Ruan, X. 2016. Saliency detection with recurrent fully convolutional networks. In *Proceedings of the European Conference on Computer Vision*, 825–841.

Wei, Y.; Wen, F.; Zhu, W.; and Sun, J. 2012. Geodesic saliency using background priors. In *Proceedings of the European Conference on Computer Vision*, 29–42.

Yan, Q.; Xu, L.; Shi, J.; and Jia, J. 2013. Hierarchical saliency detection. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, 1155–1162.

Yang, C.; Zhang, L.; Lu, H.; Ruan, X.; and Yang, M.-H. 2013. Saliency detection via graph-based manifold ranking. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, 3166–3173.

Zhang, J.; Sclaroff, S.; Lin, Z.; Shen, X.; Price, B.; and Mech, R. 2015. Minimum barrier salient object detection at 80 fps. In *Proceedings of International Conference on Computer Vision*, 1404–1412.

Zhao, R.; Ouyang, W.; Li, H.; and Wang, X. 2015. Saliency detection by multi-context deep learning. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, 1265–1274.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2921–2929.