Salient Object Detection with Semantic Priors

Tam V. Nguyen
Department of Computer Science
University of Dayton
tamnguyen@udayton.edu

Luoqi Liu
Department of ECE
National University of Singapore
liuluoqi@u.nus.edu

Abstract

Salient object detection has increasingly become a popular topic in cognitive and computational sciences, including computer vision and artificial intelligence research. In this paper, we propose integrating *semantic priors* into the salient object detection process. Our algorithm consists of three basic steps. Firstly, the explicit saliency map is obtained based on the semantic segmentation refined by the explicit saliency priors learned from the data. Next, the implicit saliency map is computed based on a trained model which maps the implicit saliency priors embedded into regional features with the saliency values. Finally, the explicit semantic map and the implicit map are adaptively fused to form a pixel-accurate saliency map which uniformly covers the objects of interest. We further evaluate the proposed framework on two challenging datasets, namely, ECSSD and HKUIS. The extensive experimental results demonstrate that our method outperforms other state-of-the-art methods.

1 Introduction

Salient object detection aims to determine the salient objects which draw the attention of humans on the input image. It has been successfully adopted in many practical scenarios, including image resizing [Goferman *et al.*, 2010], attention retargeting [Nguyen *et al.*, 2013a], dynamic captioning [Nguyen *et al.*, 2013b] and video classification [Nguyen *et al.*, 2015b]. The existing methods can be classified into biologically-inspired and learning-based approaches.

The early **biologically-inspired** approaches [Itti *et al.*, 1998; Koch and Ullman, 1985] focused on the contrast of low-level features such as orientation of edges, or direction of movement. Since human vision is sensitive to color, different approaches use local or global analysis of (color-) contrast. Local methods estimate the saliency of a particular image region based on immediate image neighborhoods, e.g., based on histogram analysis [Cheng *et al.*, 2011]. While such approaches are able to produce less blurry saliency maps, they are agnostic of global relations and structures, and they may also be more sensitive to high frequency content like image edges and noise. In a global manner, [Achanta *et al.*, 2009] achieves globally consistent results by computing color dissimilarities to the mean image color. There also exist various patch-based methods which estimate dissimilarity between image patches [Goferman *et al.*, 2010; Perazzi *et al.*, 2012]. While these algorithms are more consistent in terms of global image structures, they suffer from the involved combinatorial complexity, hence they are applicable only to relatively low resolution images, or they need to operate in spaces of reduced image dimensionality [Bruce and Tsotsos, 2005], resulting in loss of salient details and highlighting edges.

For the **learning-based** approaches, [Jiang *et al.*, 2013] trained a model to learn the mapping between regional features and saliency values. Meanwhile, [Kim *et al.*, 2014] separated the salient regions from the background by finding an optimal linear combination of color coefficients in the high-dimensional color space. However, the resulting saliency maps tend to also highlight adjacent regions of salient object(s). Additionally, there exist many efforts to study visual saliency with different cues, *i.e.*, depth matters [Lang *et al.*, 2012], audio source [Chen *et al.*, 2014], touch behavior [Ni *et al.*, 2014], and object proposals [Nguyen and Sepulveda, 2015].

Along with the advancements in the field, a new challenging question is arisen "why an object is more salient than others". This emerging question appears along with the rapid evolvement of the research field. The early datasets, *i.e.*, MSRA1000 [Achanta *et al.*, 2009], only contain images with one single object and simple background. The challenge is getting more serious when more complicated saliency datasets, ECSSD [Yan *et al.*, 2013] and HKUIS [Li and Yu, 2015] are introduced with one or multiple objects in an image with complex background. This drives us to the difference between the human fixation collection procedure and the salient object labeling process. In the former procedure, the human fixation is captured when a viewer is displayed an image for 2-5 seconds under *free-viewing* settings. Within such a short period of time, the viewer only fixates to some image locations that immediately attract his/her attention. For the latter process, a labeler is given a longer time to mark the pixels belonging to the salient object(s). In case of multiple objects appearing in the image, the labeler naturally identifies the *semantic label* of each object and then decides which object is salient. This bridges the problem of salient object detection

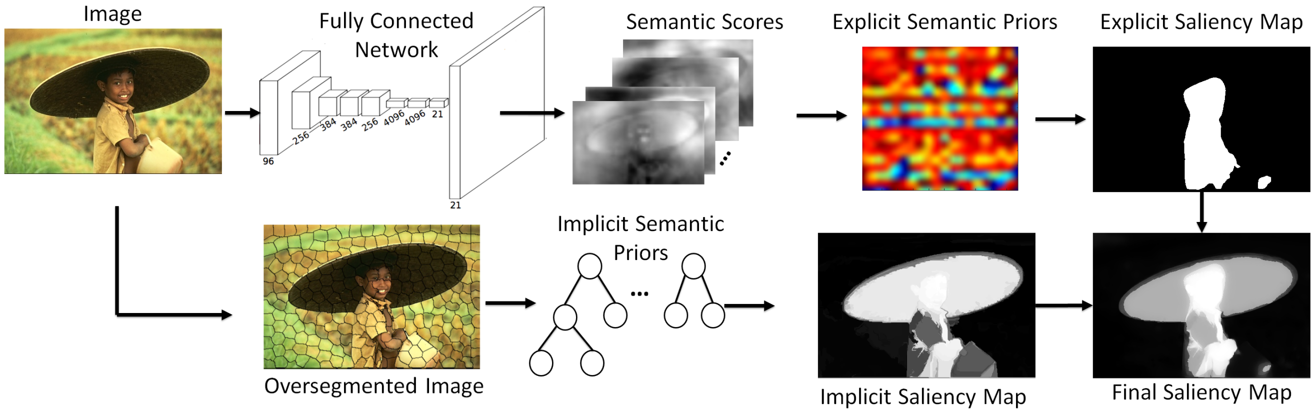


Figure 1: Pipeline of our SP saliency detection algorithm: semantic scores from the semantic extraction (Section 2.1), explicit semantic priors to compute the explicit map (Section 2.2), implicit semantic priors to compute the implicit map (Section 2.3), and saliency fusion (Section 2.4).

into the semantic segmentation research. In the latter semantic segmentation problem, the semantic label of each single pixel is decided based on a trained model which maps the features of the region containing the pixel and a particular semantic class label [Liu *et al.*, 2011]. There are many improvements in this task by handling the adaptive inference [Nguyen *et al.*, 2015a], adding object detectors [Tighe and Lazebnik, 2013], or adopting deep superpixel's features [Nguyen *et al.*, 2016]. There emerges a deep learning method, fully connected network (FCN) [Long *et al.*, 2015], which modifies the popular Convolutional Neural Networks (CNN) [Krizhevsky *et al.*, 2012] to a new deep model mapping the input pixel directly to a semantic class label. There are many works improving FCN by considering more factors such as probabilistic graphical models [Zheng *et al.*, 2015].

Recently, along with the advancement of deep learning in semantic segmentation, deep networks, such as CNN, or even FCN, have been exploited to obtain more robust features than handcrafted ones for salient object detection. In particular, **deep networks** [Wang *et al.*, 2015; Li and Yu, 2015; Li *et al.*, 2016] achieve substantially better results than previous state of the art. However, these works only focus on switching the training data (with output from semantic classes to binary classes for salient object detection problem), or adding more network layers. In fact, the impact of the semantic information is not explicitly studied in the previous deep network-based saliency models. Therefore, in this work, we investigate the application of semantic information into the problem of salient object detection. In particular, we propose the *semantic priors* to form the explicit and implicit semantic saliency maps in order to produce a high quality salient object detector. The main contributions of this work can be summarized as follows.

- We conduct the comprehensive study on how the semantic priors affect the salient object detection.
- We propose the explicit saliency map and the implicit saliency map, derived from the semantic priors, which can discover the saliency object(s).
- We extensively evaluate our proposed method on two challenging datasets in order to know the impact of our work in different settings.

2 Proposed Method

In this section, we describe the details of our proposed *semantic priors* (SP) based saliency detection, and we show how to integrate the semantic priors as well as the saliency fusion can be efficiently computed. Figure 1 illustrates the overview of our processing steps.

2.1 Semantic Extraction

Saliency detection and semantic segmentation are highly correlated but essentially different in that saliency detection aims at separating generic salient objects from background, whereas semantic segmentation focuses on distinguishing objects of different categories. As aforementioned, fully connected network (FCN) [Long *et al.*, 2015] and its variant, *i.e.*, [Zheng *et al.*, 2015] are currently the state-of-the-art methods in the semantic segmentation task. Therefore, in this paper, we consider the end-to-end deep fully connected networks into our framework. Here, "end-to-end" means that the deep networks only need to be run on the input image once to produce a complete semantic map C with the same pixel resolution as the input image. We combine outputs from the final layer and the pool4 layer, at stride 16 and pool3, at stride 8. In particular, we obtain the confidence score $C_{x,y}$ for each single pixel (x, y) as below.

$$C_{x,y} = \{C^1_{x,y}, C^2_{x,y}, \cdots, C^{n_c}_{x,y}\}, \quad (1)$$

where $C^1_{x,y}, C^2_{x,y}, \cdots, C^{n_c}_{x,y}$ indicate the likelihood that the pixel (x, y) belongs to the listed n_c semantic classes. Given an input image with size $h \times w$, the dimension of C is $h \times w \times n_c$.

2.2 Explicit Saliency Map

The objective of the explicit saliency map is to capture the preference of humans over different semantic classes. In other words, we aim to investigate which class is favoured by humans if there exist two or more classes in the input image. The class label $L_{x,y}$ of each single pixel (x, y) can be obtained as below:

$$L_{x,y} = \arg\max C_{x,y}. \quad (2)$$

Next, given a groundtruth map G, the density of each semantic class k in the input image is defined by:

$$p_k = \frac{\sum_{x,y}(L_{x,y} = k) \times G_{x,y}}{\sum_{x,y}(L_{x,y} = k)}, \quad (3)$$

where $(L_{x,y} = k)$ is a boolean expression which verifies whether $L_{x,y}$ is equivalent to class k. Note that the size of the groundtruth map is also $h \times w$. Given the training dataset, we extract the co-occurrence saliency pairwise of one class and other $n_c - 1$ classes. The pairwise value $\theta_{g,t}$ of two semantic classes g and t is computed as below.

$$\theta_{k,t} = \begin{cases} 1 & , \exists L_{x',y'} = k \wedge L_{x'',y''} = t \\ 0 & , \text{otherwise} \end{cases}. \quad (4)$$

We define the *explicit semantic priors* as the accumulated co-occurrence saliency pairwise of all classes. The explicit semantic priors of two classes g and t is calculated as below.

$$sp^{Explicit}_{k,t} = \frac{\sum_{i=1}^{n_t} p^i_k \theta^i_{k,t}}{\sum_{i=1}^{n_t} \theta^i_{k,t} + \epsilon}, \quad (5)$$

where n_t is the number of images in the training set, and ϵ is inserted to avoid the division by zero. For the testing phase, given a test image, the explicit saliency value of each single pixel (x, y) is computed as:

$$S^{Explicit}_{x,y} = \sum_{k=1}^{n_c}\sum_{t=1}^{n_c}(L_{x,y} = k) \times \theta_{k,t} \times sp^{Explicit}_{k,t}. \quad (6)$$

2.3 Implicit Saliency Map

Obviously the explicit saliency map performs well in case of the detected objects are in the listed class labels. However, the explicit saliency map fails in case of the salient objects are not in the n_c class labels. Therefore, we propose the implicit saliency map which can uncover the salient objects not in the listed semantic classes. To this end, we oversegment the input image into non-overlapping regions. Then we extract features from each image region. Different from other methods which simply learn the mapping between the locally regional features with the saliency values, here, we take the semantic information into consideration. In particular, we are interested in studying the relationship between the regional features with the saliency values under the impact of semantic-driven features. Therefore, besides the off-the-shelf region features, we add two new features for each image region, namely, global semantic and local semantic features. The local semantic feature of each image region q is defined as:

$$sp_{1,q} = \frac{\sum_{x,y} G_{x,y} \times (r(x, y) = q)}{\sum_{x,y} r(x, y) = q}, \quad (7)$$

Table 1: The regional features. Two sets of semantic features are included, namely sp_1 and sp_2.

Description	Dim
The average normalized coordinates	2
The bounding box location	4
The aspect ratio of the bounding box	1
The normalized perimeter	1
The normalized area	1
The normalized area of the neighbor regions	1
The variances of the RGB values	3
The variances of the L*a*b* values	3
The aspect ratio of the bounding box	3
The variances of the HSV values	3
Textons [Leung and Malik, 2001]	15
The local semantic features sp_1	n_c
The global semantic features sp_2	n_c

where $r(x, y)$ returns the region index of pixel (x, y). Meanwhile, the global semantic feature of the image region q is defined as:

$$sp_{2,q} = \frac{\sum_{x,y} C_{x,y} \times (r(x, y) = q)}{h \times w}. \quad (8)$$

The semantic features $sp^{Implicit} = \{sp_1, sp_2\}$ are finally combined with other regional features. We consider the semantic features here as the *implicit semantic priors* since they implicitly affect the mapping of the regional features and saliency scores. All of regional features are listed in Table 1. Then, we learn a regressor rf which maps the extracted features to the saliency values. In this work, we adopt the random forest regressor in [Jiang *et al.*, 2013] which demonstrates a good performance. The training examples include a set of n_r regions $\{\{r_1, sp^{Implicit}_1\}, \{r_2, sp^{Implicit}_2\}, \cdots, \{r_{n_r}, sp^{Implicit}_{n_r}\}\}$ and the corresponding saliency scores $\{s_1, s_2, \cdots, s_{n_r}\}$, which are collected from the oversegmentation across a set of images with the ground truth annotation of the salient objects. The saliency value of each training image region is set as follows: if the number of the pixels (in the region) belonging to the salient object or the background exceeds 80% of the number of the pixels in the region, its saliency value is set as 1 or 0, respectively.

For the testing phase, given the input image, the implicit saliency value of each image region q is computed by feeding the extracted features into the trained regressor rf:

$$S^{Implicit}_q = rf(\{r_q, sp^{Implicit}_q\}). \quad (9)$$

2.4 Saliency Fusion

Given an input image with a size $h \times w$, the saliency maps from both aforementioned saliency maps are fused at the end. In particular, we scale the implicit saliency map $S^{Implicit}$, explicit saliency map $S^{Explicit}$, to the range [0..1]. Then we combine these maps as follows to compute a saliency value S for each pixel:

$$S = \alpha S^{Explicit} + \gamma S^{Implicit}, \quad (10)$$

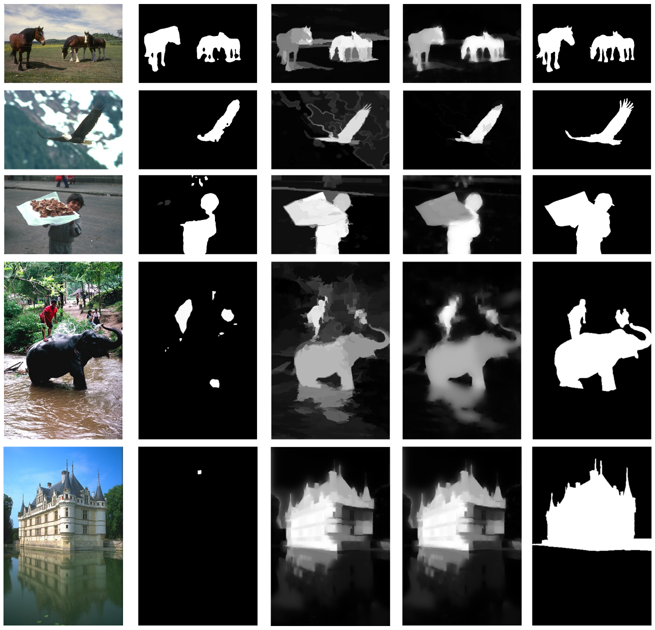


Figure 2: From left to right: the original image, the explicit saliency map, the implicit saliency map, our final saliency map, the groundtruth map. From top to bottom: in the first two rows, the explicit map helps remove the background noise from the implicit map; (third row) the implicit map recovers the food tray held by the boy; (fourth row) the elephant is revealed owing to the implicit map; (fifth row) the building is fully recovered by the implicit map. Note that the *food tray*, *elephant*, and *building* are not in the listed semantic classes of the PASCAL VOC dataset.

where the weights α is adaptively set as $\frac{\sum_{x,y} S^{Implicit}_{x,y}}{h \times w}$. Actually α measures how large the semantic pixels occupied in the image. Meanwhile, γ is set as $1-\alpha$. The resulting pixel-level saliency map may have an arbitrary scale. Therefore, in the final step, we rescale the saliency map to the range [0..1] or to contain at least 10% saliency pixels. Fig. 2 demonstrates that the two individual saliency maps, *i.e.*, explicit and implicit ones, complement each other in order to yield the good result.

2.5 Implementation Settings

For the implementation, we adopt the extension of FCN, namely CRF-FCN [Zheng *et al.*, 2015], to perform the semantic segmentation for the input image. In particular, we utilize the CRF-FCN model trained from the PASCAL VOC 2007 dataset [Everingham *et al.*, 2010] with 20 semantic classes[1]. Therefore, the regional feature's dimensionality is 79. We trained our SP framework on HKUIS dataset [Li and Yu, 2015] (training part) which contains $4,000$ pairs of images and groundtruth maps. For the image over-segmentation, we adopt the method of [Achanta *et al.*, 2012]. We set the number of regions as 200 as a trade-off between the fine over-segmentation and the processing time.

3 Evaluation

3.1 Datasets and Evaluation Metrics

We evaluate and compare the performances of our algorithm against previous baseline algorithms on two challenging benchmark datasets: ECSSD [Yan *et al.*, 2013] and HKUIS [Li and Yu, 2015] (testing part). The ECSSD dataset contains 1,000 images with the complex and cluttered background. Meanwhile, the HKUIS contains 1,447 images. Note that each image in both datasets contains single or multiple salient objects.

The first evaluation compares the precision and recall rates. In the first setting, we compare binary masks for every threshold in the range [0..255]. In the second setting, we use the image dependent adaptive threshold proposed by [Achanta *et al.*, 2009], defined as twice the mean value of the saliency map S. In addition to precision and recall we compute their weighted harmonic mean measure or $F-measure$, which is defined as:

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}. \tag{11}$$

As in previous methods [Achanta *et al.*, 2009; Perazzi *et al.*, 2012], we use β^2 = 0.3.

For the second evaluation, we follow [Perazzi *et al.*, 2012] to evaluate the mean absolute error (MAE) between a continuous saliency map S and the binary ground truth G for all image pixels (x, y), defined as:

$$MAE = \frac{1}{h \times w} \sum_{x,y} |S_{x,y} - G_{x,y}|. \tag{12}$$

3.2 Performance on ECSSD dataset

For the evaluation, we compare our work with 17 state-of-the-art methods by running the approaches' publicly available source code: attention based on information maximization (AIM [Bruce and Tsotsos, 2005]), boolean map saliency (BMS [Zhang and Sclaroff, 2016]), context-aware saliency (CA [Goferman *et al.*, 2010]), discriminative regional feature integration (DRFI [Jiang *et al.*, 2013]), frequency-tuned saliency (FT [Achanta *et al.*, 2009]),global contrast saliency (HC and RC [Cheng *et al.*, 2011]), high-dimensional color transform (HDCT [Kim *et al.*, 2014]), hierarchical saliency (HS [Yan *et al.*, 2013]), spatial temporal cues (LC [Zhai and Shah, 2006]), local estimation and global search (LEGS [Wang *et al.*, 2015]), multiscale deep features (MDF [Li and Yu, 2015]), multi-task deep saliency (MTDS [Li *et al.*, 2016]), principal component analysis (PCA [Margolin *et al.*, 2013]), saliency filters (SF [Perazzi *et al.*, 2012]), induction model (SIM [Murray *et al.*, 2011]), saliency using natural statistics (SUN [Zhang *et al.*, 2008]. Note that LEGS, MDF, and MTDS are deep learning based methods. The visual comparison of saliency maps generated from our method and different baselines are demonstrated in Figure 3. Our results are close to ground truth and focus on the main salient objects. As shown in Figure 4a,b , our work

[1] There are 20 semantic classes in the PASCAL VOC 2007 ('aeroplane', 'bicycle', 'bird', 'boat', 'bottle', 'bus', 'car', 'cat', 'chair', 'cow', 'diningtable', 'dog', 'horse', 'motorbike', 'person', 'pottedplant', 'sheep', 'sofa', 'train', 'tvmonitor'); and an extra 'others' class label.

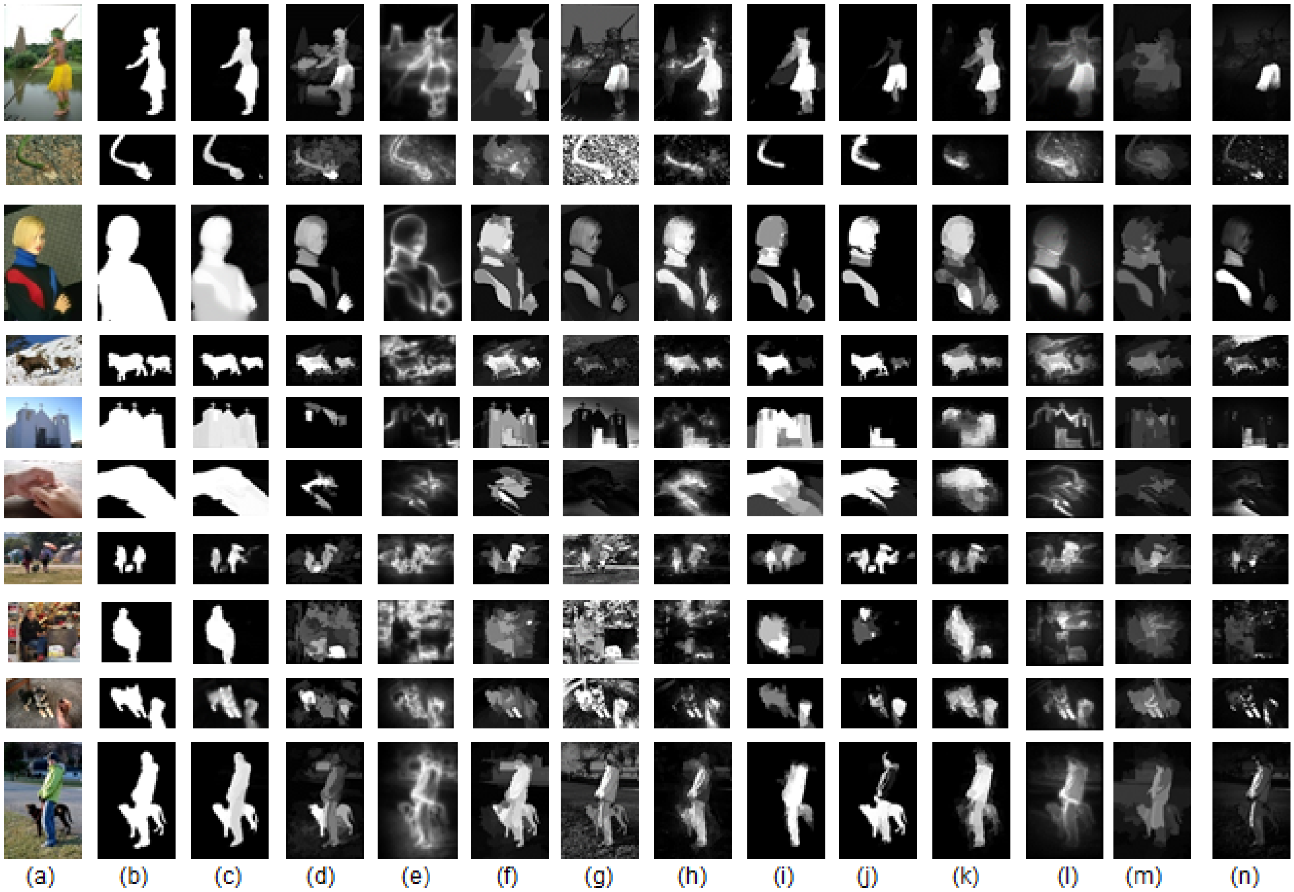


Figure 3: Visual comparison of saliency maps. From left to right: (a) Original images, (b) ground truth, (c) our SP method, (d) BMS, (e) CA, (f) DRFI, (g) FT , (h) HDCT, (i) LEGS, (j) MDF, (k) MTDS, (l) PCA, (m) RC, (n) SF. Most results are of low resolution or highlight edges whereas our final result focuses on the main salient object as shown in ground truth map (c).

reaches the highest precision/recall rate over all baselines. As a result, our method also obtains the best performance in terms of F-measure.

As discussed in [Perazzi *et al.*, 2012], neither the precision nor recall measure considers the true negative counts. These measures favor methods which successfully assign saliency to salient pixels but fail to detect non-salient regions over methods that successfully do the opposite. Thus they suggested that MAE is a better metric than precision recall analysis for this problem. As shown in Figure 4c, our work outperforms the state-of-the-art performance [Li and Yu, 2015] by 10%.

3.3 Performance on HKUIS dataset

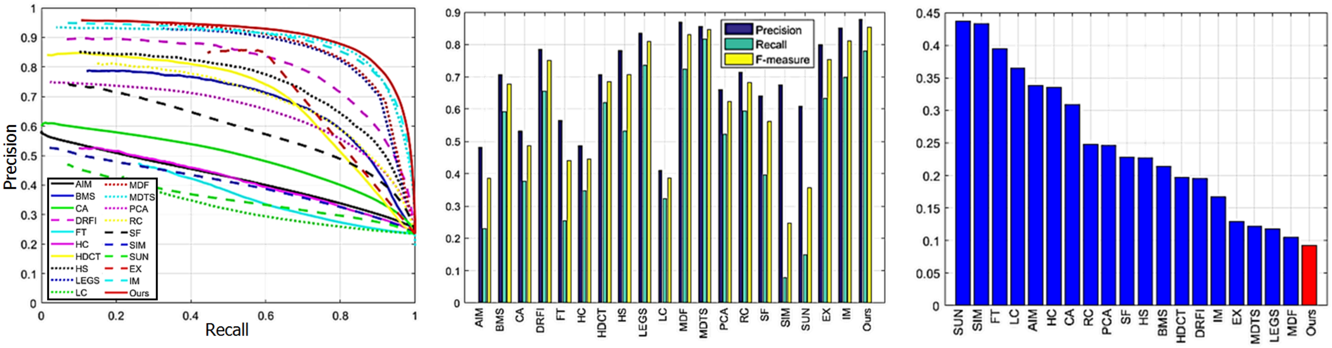
Since HKUIS is a relatively new dataset, we only have 15 baselines. We first evaluate our methods using a precision/recall curve which is shown in Figure 5a, b. Our method outperforms all other baselines in both two settings, namely fixed threshold and adaptive threshold. As shown in Figure 5c, our method achieves the best performance in terms of MAE. In particular, our work outperforms other methods by a large margin, 25%.

3.4 Effectiveness of Explicit and Implicit Saliency Maps

We also evaluate the individual components in our system, namely, the explicit saliency map (EX), and the implicit saliency map (IM), in both ECSSD and HKUIS. As shown in Fig. 4 and Fig. 5, the two components generally achieve the acceptable performance (in terms of precision, recall, F-measure and MAE) which is comparable to other baselines. EX outperforms IM in terms of MAE, whereas IM achieves a better performance in terms of F-measure. When adaptively fusing them together, our unified framework achieves the state-of-the-art performance in every single evaluation metric. That demonstrates that these individual components complement each other in order to yield the good result.

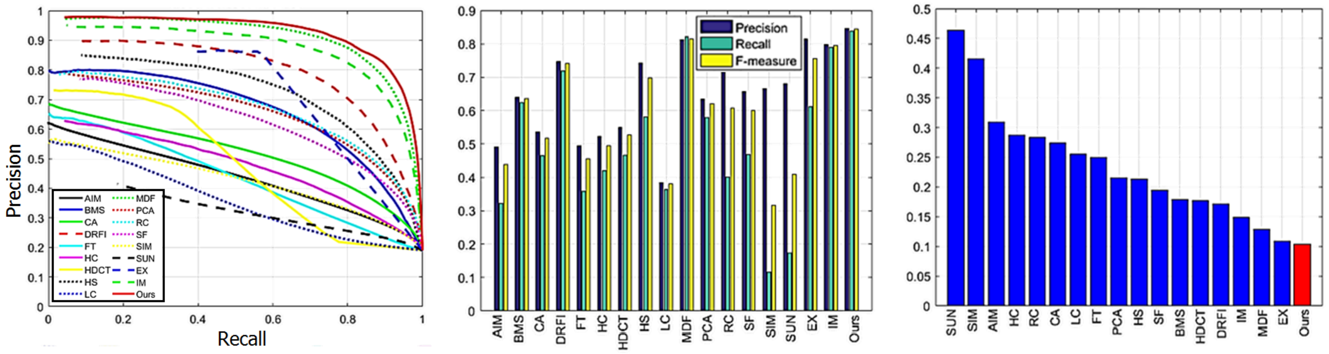
3.5 Computational Efficiency

It is also worth investigating the computational efficiency of different methods. In Table 2, we compare the average running time for a typical 300 × 400 image of our approach to other methods. The average time is taken on a PC with Intel i7 2.6 GHz CPU and 8GB RAM with our unoptimized Matlab code. Performance of all the methods compared in this table are based on implementations in C++ and Matlab. Basi-



(a) Fixed threshold (b) Adaptive threshold (c) Mean absolute error

Figure 4: Statistical comparison with 17 saliency detection methods using all the 1, 000 images from ECSSD dataset [Yan *et al.*, 2013] with pixel accuracy saliency region annotation: (a) the average precision recall curve by segmenting saliency maps using fixed thresholds, (b) the average precision recall by adaptive thresholding (using the same method as in FT [Achanta *et al.*, 2009], SF [Perazzi *et al.*, 2012], etc.), (c) the mean absolute error of the different saliency methods to ground truth mask.



(a) Fixed threshold (b) Adaptive threshold (c) Mean absolute error

Figure 5: Statistical comparison with 15 saliency detection methods using all the 1, 447 images from the test set of HKUIS benchmark [Li and Yu, 2015] with pixel accuracy saliency region annotation: (a) the average precision recall curve by segmenting saliency maps using fixed thresholds, (b) the average precision recall by adaptive thresholding (using the same method as in FT [Achanta *et al.*, 2009], etc.), (c) the mean absolute error of the different saliency methods to ground truth mask.

Table 2: Runtime comparison of different methods.

Method	CA	DRFI	SF	RC	Ours
Time (s)	51.2	10.0	0.15	0.25	3.8
Code	Matlab	Matlab	C++	C++	Matlab

cally, C++ implementation runs faster than the Matlab based code. The CA method [Goferman *et al.*, 2010] is the slowest one because it requires an exhaustive nearest-neighbor search among patches. Meanwhile, our method is able to run faster than other Matlab based implementations. Our procedure spends most of the computation time on semantic segmentation and extracting regional features.

4 Conclusion and Future Work

In this paper, we have presented a novel method, *semantic priors* (SP), which adopts the semantic segmentation in order to detect *salient objects*. To this end, two maps are derived from semantic priors: the explicit saliency map and the implicit saliency map. These two maps are fused together to give a saliency map of the salient objects with sharp boundaries. Experimental results on two challenging datasets demonstrate that our salient object detection results are 10% - 25% better than the previous best results (compared against 15+ methods in two different datasets), in terms of mean absolute error.

For future work, we aim to investigate other sophisticated techniques for semantic segmentation with a larger number of semantic classes. Also, we would like to study the reverse impact of salient object detection into the semantic segmentation process.

Acknowledgements

Thanh Duc Ngo, Khang Nguyen from Multimedia Communications Laboratory, University of Information Technology are additional authors. We gratefully acknowledge the support of NVIDIA Corporation with the GPU donation.

References

[Achanta *et al.*, 2009] Radhakrishna Achanta, Sheila S. Hemami, Francisco J. Estrada, and Sabine Süsstrunk. Frequency-tuned salient region detection. In *CVPR*, pages 1597–1604, 2009.

[Achanta *et al.*, 2012] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurélien Lucchi, Pascal Fua, and Sabine Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE T-PAMI*, 34(11):2274–2282, 2012.

[Bruce and Tsotsos, 2005] Neil D. B. Bruce and John K. Tsotsos. Saliency based on information maximization. In *NIPS*, 2005.

[Chen *et al.*, 2014] Yanxiang Chen, Tam V. Nguyen, Mohan S. Kankanhalli, Jun Yuan, Shuicheng Yan, and Meng Wang. Audio matters in visual attention. *T-CSVT*, 24(11):1992–2003, 2014.

[Cheng *et al.*, 2011] Ming-Ming Cheng, Guo-Xin Zhang, Niloy J. Mitra, Xiaolei Huang, and Shi-Min Hu. Global contrast based salient region detection. In *CVPR*, pages 409–416, 2011.

[Everingham *et al.*, 2010] Mark Everingham, Luc J. Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010.

[Goferman *et al.*, 2010] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. In *CVPR*, pages 2376–2383, 2010.

[Itti *et al.*, 1998] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE T-PAMI*, 20(11):1254–1259, 1998.

[Jiang *et al.*, 2013] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. Salient object detection: A discriminative regional feature integration approach. In *CVPR*, pages 2083–2090, 2013.

[Kim *et al.*, 2014] Jiwhan Kim, Dongyoon Han, Yu-Wing Tai, and Junmo Kim. Salient region detection via high-dimensional color transform. In *CVPR*, pages 883–890, 2014.

[Koch and Ullman, 1985] C Koch and S Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol*, 1985.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012.

[Lang *et al.*, 2012] Congyan Lang, Tam V. Nguyen, Harish Katti, Karthik Yadati, Mohan S. Kankanhalli, and Shuicheng Yan. Depth matters: Influence of depth cues on visual saliency. In *ECCV*, pages 101–115, 2012.

[Leung and Malik, 2001] Thomas K. Leung and Jitendra Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV*, 43(1):29–44, 2001.

[Li and Yu, 2015] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *CVPR*, pages 5455–5463, 2015.

[Li *et al.*, 2016] Xi Li, Liming Zhao, Lina Wei, Ming-Hsuan Yang, Fei Wu, Yueting Zhuang, Haibin Ling, and Jingdong Wang. Deepsaliency: Multi-task deep neural network model for salient object detection. *IEEE T-IP*, 25(8):3919–3930, 2016.

[Liu *et al.*, 2011] Ce Liu, Jenny Yuen, and Antonio Torralba. Nonparametric scene parsing via label transfer. *IEEE T-PAMI*, 33(12):2368–2382, 2011.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[Margolin *et al.*, 2013] Ran Margolin, Ayellet Tal, and Lihi Zelnik-Manor. What makes a patch distinct? In *CVPR*, pages 1139–1146, 2013.

[Murray *et al.*, 2011] Naila Murray, Maria Vanrell, Xavier Otazu, and C. Alejandro Párraga. Saliency estimation using a nonparametric low-level vision model. In *CVPR*, pages 433–440, 2011.

[Nguyen and Sepulveda, 2015] Tam V. Nguyen and Jose Sepulveda. Salient object detection via augmented hypotheses. In *IJCAI*, pages 2176–2182, 2015.

[Nguyen *et al.*, 2013a] Tam V. Nguyen, Bingbing Ni, Hairong Liu, Wei Xia, Jiebo Luo, Mohan S. Kankanhalli, and Shuicheng Yan. Image re-attentionizing. *IEEE T-MM*, 15(8):1910–1919, 2013.

[Nguyen *et al.*, 2013b] Tam V. Nguyen, Mengdi Xu, Guangyu Gao, Mohan S. Kankanhalli, Qi Tian, and Shuicheng Yan. Static saliency vs. dynamic saliency: a comparative study. In *ACM Multimedia*, pages 987–996, 2013.

[Nguyen *et al.*, 2015a] Tam V. Nguyen, Canyi Lu, Jose Sepulveda, and Shuicheng Yan. Adaptive nonparametric image parsing. *T-CSVT*, 25(10):1565–1575, 2015.

[Nguyen *et al.*, 2015b] Tam V. Nguyen, Zheng Song, and Shuicheng Yan. STAP: spatial-temporal attention-aware pooling for action recognition. *T-CSVT*, 25(1):77–86, 2015.

[Nguyen *et al.*, 2016] Tam V. Nguyen, Luoqi Liu, and Khang Nguyen. Exploiting generic multi-level convolutional neural networks for scene understanding. In *ICARCV*, pages 1–6, 2016.

[Ni *et al.*, 2014] Bingbing Ni, Mengdi Xu, Tam V. Nguyen, Meng Wang, Congyan Lang, ZhongYang Huang, and Shuicheng Yan. Touch saliency: Characteristics and prediction. *IEEE T-MM*, 16(6):1779–1791, 2014.

[Perazzi *et al.*, 2012] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, pages 733–740, 2012.

[Tighe and Lazebnik, 2013] Joseph Tighe and Svetlana Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *CVPR*, pages 3001–3008, 2013.

[Wang *et al.*, 2015] Lijun Wang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Deep networks for saliency detection via local estimation and global search. In *CVPR*, pages 3183–3192, 2015.

[Yan *et al.*, 2013] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *CVPR*, pages 1155–1162, 2013.

[Zhai and Shah, 2006] Yun Zhai and Mubarak Shah. Visual attention detection in video sequences using spatiotemporal cues. In *ACM Multimedia*, pages 815–824, 2006.

[Zhang and Sclaroff, 2016] Jianming Zhang and Stan Sclaroff. Exploiting surroundedness for saliency detection: A boolean map approach. *IEEE T-PAMI*, 38(5):889–902, 2016.

[Zhang *et al.*, 2008] Lingyun Zhang, Matthew H. Tong, Tim K. Marks, Honghao Shan, and Garrison W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7), 2008.

[Zheng *et al.*, 2015] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H. S. Torr. Conditional random fields as recurrent neural networks. In *ICCV*, pages 1529–1537, 2015.