Salient Object Detection using a Context-Aware Refinement Network

Md Amirul Islam
amirul@cs.umanitoba.ca
Mahmoud Kalash
kalashm@cs.umanitoba.ca
Mrigank Rochan
mrochan@cs.umanitoba.ca
Neil D. B. Bruce
bruce@cs.umanitoba.ca
Yang Wang
ywang@cs.umanitoba.ca

Department of Computer Science
University of Manitoba
Winnipeg, MB, Canada

Abstract

Recently there has been remarkable success in pushing the state of the art in salient object detection. Most of the improvements are driven by employing end-to-end deeper feed-forward networks. However, in many cases precisely detecting salient regions requires representation of fine details. Combining high-level and low-level features using skip connections is a strategy that has been proposed, but sometimes fails to select the right contextual features. To overcome this limitation, we propose an end-to-end encoder-decoder network that employs recurrent refinement to generate a saliency map in a coarse-to-fine fashion by incorporating finer details in the detection framework. The proposed approach makes use of refinement units within each stage of the decoder that are responsible for refining the saliency map produced by earlier layers by learning context-aware features. Experimental results on several challenging saliency detection benchmarks validate the effectiveness of our proposed architecture providing a significant improvement over current state-of-the-art methods.

1 Introduction

Salient object detection aims to precisely detect objects that capture human attention in images, or that are the main subject of the image. In recent years, there have been significant advances in developing models for salient object detection that have achieved a great deal of success, motivated by a wide rage of applications (e.g. semantic segmentation, object detection, image summarization, content-aware image cropping and others).

Traditional saliency detection methods such as DRFI [9], DSR [17], HS [31] mostly focus on relatively general cues like contrast, color, texture that tend to be diagnostic of what is salient to evaluate the distinctiveness of each image region or pixel considering local and global contextual information. In most cases, this class of methods tries to highlight

Image GT Initial Saliency Map Refined Saliency Map After CRF

Figure 1: An example of applying context-aware refinement network to an initial saliency map produced by the encoder network. Compared to initial saliency map, the refined saliency map has significantly sharper edges and better spatial information.

object boundaries or interiors uniformly but in many cases fails to preserve object details. Moreover, some of the classic approaches are often unable to detect salient objects with large sizes, and targets of a photograph that lie on complex textures and backgrounds.

Recent success of Convolutional Neural Networks (CNN) in a variety of computer vision tasks (e.g. image classification [3, 26], semantic segmentation [1, 21], edge detection [29]) has attracted wide attention to these methods, and has motivated efforts towards using Fully Convolutional Networks (FCN) for the salient object detection [4, 12, 14, 19, 28, 33] task. One popular solution is the use of an encoder-decoder based approach that applies stage-wise refinement to capture finer details resident in early convolutional layers through skip connections at the decoding stage. However, simply integrating features of higher spatial dimension within the refinement (decoding) process does not always achieve significant improvements. Inspired by the success of encoder-decoder networks in pixel-wise labeling tasks, we apply a network with this structure to detect salient regions in an end-to-end fashion. This takes the form of our proposed context-aware refinement network wherein the decoder part takes coarse saliency maps generated by the encoder network and hierarchically refines the saliency map to produce a final output that matches the resolution of the input. To overcome the limitations with existing approaches, we propose a refinement unit that takes full advantage of the high spatial dimension features from earlier layers in the refinement process. As demonstrated in Fig. 1, we observe that high-level features can better locate the salient object and low-level features capture rich spatial information. With that being said, we believe that integrating high-level features with low-level features is useful in the salient object detection task. In this paper, we propose a new approach for salient object detection inspired by the previous approaches. Our contributions can be summarized as follows:

- We propose a novel end-to-end encoder-decoder based salient object detection model that can simultaneously predict saliency maps at different resolutions.
- We propose a context-aware refinement network, which serves as the decoder network, and can hierarchically and progressively refine saliency maps to recover fine details of the image by integrating local and global contextual information through gate units, global convolution units and boundary refinement units. Moreover, we combine the prior map with the final prediction map. Furthermore, our model is general enough that it can be easily applied to other pixel-wise labeling tasks (e.g. semantic segmentation, scene labeling, depth estimation etc.).
- Experimental results on four benchmark datasets and comparisons with recent state-of-the-art approaches demonstrate the effectiveness and superiority of our approach on the salient object detection task.

2 Related Work

Over the past few years, a large number of salient object detection approaches have been developed. In general, those approaches can be classified into two main categories, i.e., either contrast-based methods that use hand-crafted features or methods that apply deep learning to learn both features and the classifier.

Contrast based methods select and combine important features to detect objects that attract attention. Some of these methods use local, low-level features such as multi-scale color, intensity and orientation filters [7], mid-level visual cues [30], or the contrast of multiple feature distributions [10]. However, other methods use global features like region descriptors [9], global region contrast [2], or a combination of features (i.e. multi-scale contrast, center surround histograms, and color spatial distributions) [20].

More recently, CNN have shown superior performance compared to these traditional methods on commonly used benchmarks. CNN based models have raised the bar on the quality of predictions possible in multiple fields of computer vision, including salient object detection. Recently, many salient object detection methods adopt CNN based models due to the ability to extract more representative and complex high-level features. Li and Yu [13] proposed a deep neural network that extracts features from three differently scaled input maps and then aggregates them into one saliency map. Wang et al. [27] integrated both local estimation and global search using two sequential CNN to predict saliency maps. Local saliency information (i.e. the saliency value for each pixel) is extracted by the first CNN and then forwarded along with the global contrast and geometric information to the second CNN for further refinement. Zhao et al. [33] proposed a multi-context CNN that obtains and integrates global and local context information to produce saliency maps. Liu and Han [19] tackled the salient object detection problem in a global to local (coarse to fine) manner. Their architecture follows the end-to-end encoder-decoder approach where the encoder learns multiple global structured saliency cues and their optimal combination to produce a coarse saliency map. Then, another hierarchical recurrent convolutional neural network refines the coarse saliency map stage-by-stage by integrating local contextual information. Li and Yu [14] proposed an end-to-end deep contrast network with two streams to enhance salient object boundary detection. They combine a pixel-level fully convolutional stream that produces a saliency map with pixel-level accuracy and a segment-wise spatial pooling stream that extracts segment-wise features. The fused saliency map is finally refined with a fully connected CRF model. Wang et al. [28] proposed a recurrent fully convolutional network for saliency detection. In the first time step, they use the potential salient regions in the input image as a prior knowledge of possibly salient regions in order to predict an initial saliency map. This in turn serves as the saliency prior map for the next time step.

In contrast to above approaches, we perform a step-by-step multi-stage supervised refinement for the encoded saliency map until the saliency map spatial resolution matches the input image. This also notably includes specific mechanisms for how early feature information is routed in making a final determination of saliency.

3 Context-Aware Refinement Network

In this section, we discuss our proposed context-aware refinement network to address the problem of salient object detection. Then, we design a fully-convolutional feedback refinement network using context-aware refinement units.

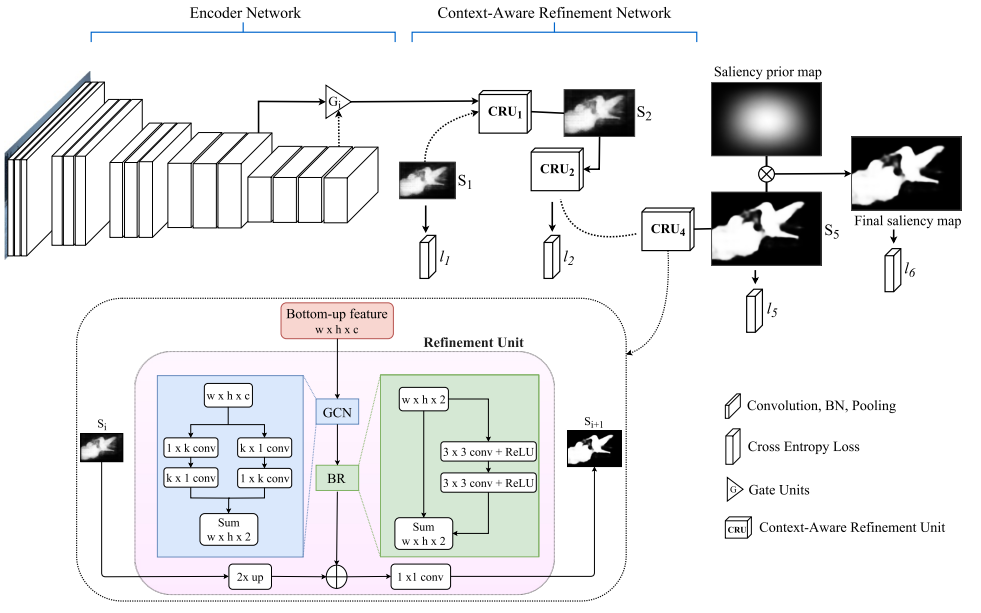


Figure 2: An overview of the proposed salient object detection framework. The bottom-up (encoder) network consisting of multiple layers (e.g. convolution, batch normalization, pooling, ReLU) is integrated with the refinement network through skip connections. The refinement network has context-aware refinement units (CRU_1, CRU_2,..., CRU_4) that take a bottom-up feature map and previous stage prediction map (S_i) to generate subsequent stage prediction maps (S_{i+1}). We down-sampled the ground-truth saliency maps to incorporate stage-wise supervision (l_1, l_2,..., l_6) in the refinement network. GCN and BR are used in CRUs (see Sec. 3.2 for details). Note that we also combine the final prediction map with the saliency prior to obtain the final saliency map.

3.1 Overview

We adopt the popular encoder-decoder network architecture for salient object detection, where a CNN initially encodes the input image to produce a coarse spatial resolution prediction, and then a refinement network decodes the coarse saliency map to provide a full resolution pixel-wise prediction map. Our overall salient object detection network is illustrated in Fig. 2. We employ pre-trained ResNet-101 [3] as the encoder network, and we propose a novel refinement network that serves as our decoder network. We extract multi-scale feature maps from different stages of the encoder. The context-aware refinement network uses these feature maps to generate semantic score maps in each stage of the refinement network. Similar to previous approaches [5, 19, 24], semantic score maps for lower resolutions are upsampled through bilinear interpolation and combined with the feature maps from the encoder to generate a higher resolution refined saliency map. In our case, this combination is influenced by the refinement units involved. The semantic score map generated from the last stage of the refinement unit is treated as the final prediction map of our network. In addition, the saliency prior map is integrated with the final prediction map as a final stage of refinement before evaluating the loss function. In the following section, we discuss the context-aware refinement unit and saliency prior map.

3.2 Context-Aware Refinement Unit

The task of salient object detection requires per-pixel classification as well as correct localization. Current state-of-the-art salient object detection methods [4, 19] mostly target the design principles of the decoding process such that an initial prediction map is refined to produce the full resolution map. However, in most cases, refinement is done across different levels by combining convolution features from the encoder (or feature extractor) network with decoder layers. Directly combining convolution features with the semantic score map through concatenation or element-wise summation may have unpredictable consequences, and has the possibility of degrading the contribution of lower depth feature maps (semantic score maps). Hence, in only using features from the encoder network through skip connections [5, 21] there are inherent limits on spatial detail that may be recovered since the model cannot take full advantage of the higher resolution feature maps. Therefore, following previous work [6], we integrate a multiplicative gate unit at each stage of refinement that controls the information being passed forward to resolve ambiguity between background and salient object classes.

Moreover, in salient object detection, the object is often biased in its position towards the center of the image and the classifier has a view of the entire object in context only within the deepest layers of the encoder. However, if the salient object is resized to a large scale, then the receptive field (kernel) of the skip connections covers only a part of the object, which can be problematic in refining missing details. In [6], all the skip connections in gate units and refinement units use a 3×3 receptive field to generate semantic score maps.

Based on the above observations, and also drawing inspiration from [23], we design a refinement unit that is mainly composed of a Global Convolution Network (GCN) and a Boundary Refinement (BR) block that overcomes these drawbacks. GCN uses a combination of $1 \times k + k \times 1$ and $k \times 1 + 1 \times k$ convolutions, resulting in a $k \times k$ convolution that enables dense connection within the $k \times k$ region (instead of directly using a $k \times k$ kernel), and thereby helping to capture broader context. BR consists of stack of two 3×3 convolutions followed by element-wise summation to further refine the boundary pixels. We now describe the detailed architecture of context-aware refinement units (CRU).

The detailed architecture of the CRU is illustrated in the dotted box of Fig. 2 which has two input paths. Our refinement units are generic and can be modified to accept an arbitrary number of feature maps with different resolutions. Note that, although these units are identical, they do not share weights among them since each unit learns to recover missing spatial information in order to resolve ambiguity during refinement stages. It is also noteworthy that each CRU combines feature representations obtained from different levels of the encoder network.

The first input of each refinement unit is comprised of a bottom-up feature map derived from a multiplicative gate unit that serves a primary role of filtering out ambiguity between background and salient objects by controlling the activation from features passed from encoder layers to decoder layers. The saliency map predicted from the prior stage S_i serves as the second input to the refinement unit. To that end, the first input is passed sequentially through a global convolution unit and boundary refinement unit before being combined with the 2x upsampled saliency map derived from the prior stage through concatenation followed by a 1×1 convolution across layers. The formulation of getting a bottom-up feature map from a gate unit is described by the following equations:

$$v_i = T_f(C^{i+1}), \quad u_i = T_f(C^i), \quad Z_f^i = v_i \odot u_i \tag{1}$$

where $\odot$ denotes an element-wise product. Note that, C^i and C^{i+1} are the feature map from i_{th} and $(i+1)_{th}$ layer in the encoder which are passed through a transformation function T_f to map these to semantic score maps. As noted earlier, Z_f^i is then fed to the refinement unit as the first input.

In summary, the refinement unit at each stage takes the bottom-up feature Z_f^i and last stage prediction map S_i as inputs and generates the next stage prediction map S_{i+1} through the series of operations mentioned earlier. The operations inside each refinement unit are as follows:

$$S_{i+1} = \mathbb{C}_{1\times1}(\rho(\phi(Z_f^i)) \oplus \mathbb{U}(S_i)) \tag{2}$$

where $\mathbb{C}$, ρ, ϕ, $\oplus$, and $\mathbb{U}$ denotes 1×1 convolution, global convolution unit, boundary refinement unit, concatenation, and 2x upsample operation respectively.

3.3 Saliency Prior Map

We also integrate a saliency prior map as an additional input to the network. We first calculate the per-pixel average of ground-truth training images which serves as the saliency prior map for the network. If pixels marked salient were uniformly distributed, this prior would have no effect. However, salient objects tend to be near the center of the image (likely due to compositional bias) in a manner determined by the purpose of dataset and how it was composed. Taking this into consideration, it is important to model such spatial bias and we do so by creating a prior distribution S_p that is multiplied element-wise with the final predicted saliency map S'. We convolve the final prediction layer with a Gaussian G_σ to regularize the predictions. Since the final prediction layer has two output channels (foreground and background), we slice the feature map to separate them. Note that only foreground feature slices are combined with the prior map since these contain the objectness score for each pixel that corresponds roughly to different semantic categories. We summarize the operations as follows:

$$S_p(i,j) = \frac{1}{N}\sum_{m=1}^{N}\sum_{i=1}^{h}\sum_{j=1}^{w} S_m(i,j) \qquad S_i' = S_i \times G_\sigma \qquad S_i'' = S' \odot S_p \tag{3}$$

3.4 Training with Multi-stage Supervision

Inspired by [5, 11, 19], we apply multi-stage supervision in our end-to-end network. More specifically, assume $I_m \in \mathbb{R}^{h\times w\times d}$ to be a training instance with ground-truth saliency mask $S_m \in \mathbb{R}^{h\times w}$. We obtain m resized ground-truth maps $(R_1, R_2, ..., R_m)$ by resizing S_m. A loss function ϕ_i (pixel-wise cross entropy loss) is defined to measure the quality of predicted saliency map against the resized ground-truth saliency mask $R_i(S_m)$ at different stages of the refinement network. We can write these operations as follows:

$$\ell = \sum_{m=1}^{5} l_m \qquad l_m = \xi\left(R_i, S_{m_i}\right) \qquad \xi = \frac{1}{N}\sum_i p\log(x_i, y_i|I_i) \tag{4}$$

where ξ denotes cross-entropy loss at each stage. The final loss ℓ is the summation of cross-entropy losses across different stages of the refinement network. We train the network end-to-end using back-propagation to optimize the final loss.

4 Experiments and Results

To demonstrate the effectiveness of each component of our network architecture, and study the performance of our proposed approach, we present results from experiments on four salient object detection benchmark datasets and show quantitative and qualitative comparisons of our methods with recent state-of-the-art methods. In the following section, we firstly describe the implementation specific details. Then, we report performance on several saliency detection benchmarks followed by analysis of different variants of our approach.

4.1 Implementation Details

Our network is implemented based on the publicly available Caffe library [8]. We use a single GTX Titan X GPU for both training and testing. Inspired by [1], we use the "poly" learning rate policy. Taking training efficiency into consideration, the mini-batch size is set to 1, and loss is updated after every 10 iterations (i.e. each image is used 10 times for training). We train the network using stochastic gradient descent with momentum of 0.9, and weight decay of 0.0005. The total number of iterations is set to 20k. The weights of all the newly added convolution layers in the refinement network are randomly initialized from a standard normal distribution. Since we use the pre-trained ResNet-101 model for initializing the encoder part of our network, we normalize the data using the mean and standard deviation from VGG-16. We use pixel-wise cross entropy loss to optimize the network. As commonly done in the training procedure (due to hardware constraints), we perform random cropping of the images. During training, crop size is set to 321×321. Since all the proposed modules in our network can handle input images of different sizes, we test our network with the full resolution image. To show the effectiveness of our method and the merit of individual components, we carry out comprehensive experiments including ablation studies. We report performance for the following variants of our model including the baseline:

G-FRNet: Gated Feedback Refinement Network [6] that includes the gating mechanism prior to passing information to the refinement units. We consider G-FRNet as our base model and report its experimental results.

CARNet: Context-Aware Feedback Refinement Network built on top of G-FRNet [6] for salient object detection. We integrate the prior map within the training procedure.

CARNet†: This is the same as CARNet except that we add the global convolution network (GCN) and boundary refinement (BR) block within the refinement process.

4.2 Datasets and Evaluation Metrics

Datasets: We follow the training protocol suggested in [19]. More specifically, we use the MSRA-10K [25] dataset for training and evaluating our proposed method on four saliency detection benchmark datasets, including PASCAL-S [18], ECSSD [31], HKU-IS [13], and DUT-OMRON [32]. MSRA-10K dataset consists of 10,000 images with pixel-wise annotation for salient objects. PASCAL-S dataset contains 850 images with multiple complex objects derived from PASCAL VOC 2012 validation set that provides saliency estimates in the [0, 1] range. As suggested by the author of this dataset, we threshold the saliency values using a threshold of 0.5 to obtain the binary object mask. HKU-IS dataset provides 4,447 complex images with low-contrast and multiple salient objects in each image. Similarly, DUT-OMRON is a large dataset which contains 5,168 challenging images (one or more salient objects) with complex backgrounds.

*	ECSSD [61]		HKU-IS [22]		PASCAL-S [18]		DUT-OMRON [62]	
	F-measure	MAE	F-measure	MAE	F-measure	MAE	F-measure	MAE
RC [9]	0.741	0.187	0.726	0.165	0.640	0.225	-	-
DSR [17]	0.737	0.173	0.735	0.140	0.646	0.204	-	-
DRFI [9]	0.787	0.166	0.783	0.143	0.679	0.221	0.664	0.150
MDF [13]	0.833	0.108	0.860	0.129	0.764	0.145	0.640	0.092
CHM [15]	0.722	0.195	0.728	0.158	0.631	0.222	-	-
MC [63]	0.822	0.107	0.781	0.098	0.721	0.147	0.703	0.088
ELD [12]	0.865	0.981	0.844	0.071	0.767	0.121	0.719	0.091
RFCN [28]	0.898	0.097	0.895	0.079	0.827	0.118	0.747	0.095
DHSNet [19]	0.905	0.061	0.892	0.052	0.820	0.091	0.740	-
DCL [14]	0.898	0.071	0.907	0.048	0.822	0.108	0.757	0.080
DSS [4]	0.915	0.052	0.913	0.039	0.830	0.080	-	-
CARNet†	0.9250	0.040	0.912	0.034	0.8341	0.086	0.7895	0.061

Table 1: Quantitative comparison (in terms of average F_β and MAE) with state-of-the-the methods. Best and second best scores are shown in red and blue text respectively.

Evaluation Metrics: Following previous work [19], we use four different standard metrics to measure the performance including precision-recall (PR) curves, F-measure, mean absolute error (MAE), and area under ROC curve (AUC). We calculate the precision and recall curve by thresholding the predicted saliency map using a set of thresholds, and compare the predicted binary map with the ground-truth map. MAE is the average pixel-wise difference between the predicted saliency map and the binary ground-truth map. We set β^2 in F-measure to 0.3 following previous works.

4.3 Performance Comparison with State-of-the-art Methods

We compare our proposed salient object detection model with state-of-the-art methods, including DSS [4], RFCN [28], DCL [14], DHS [19], MTDS [16], DRFI [9], LEGS [27], MDF [13]. The first few approaches are recent deep learning methods. Initially, we compare our approach with existing methods in terms of F-measure and MAE scores as shown in Table 1. Our approach achieves the best performance for most of the datasets. Our proposed approach is capable of not only detecting salient objects of different scales but also generating precise saliency maps in challenging scenarios (see Fig. 3). Fig. 4 presents the comparison of our method with state-of-the-art methods through PR-curves, F-measure and AUC metrics. It is clear from Fig. 4 and Table 1 that our proposed approach outperforms the existing methods with a reasonable margin.

4.4 Comparison with Different Variants

To demonstrate in greater detail the role of different components of our proposed network, we report performance of different variants (Sec. 4.1) of our network in Table 2. GFRNet is our base model, whereas CARNet is G-FRNet with spatial prior information. CARNet performs better than GFRNet due to the fact that adding prior to the network refines expectation based on interaction between spatial position and features, and thus helps providing a more precise final prediction. To further improve the performance of CARNet, we integrate GCN and BR within CARNet (i.e. CARNet†). Our final model CARNet† achieves the best performance and this gain in performance can be attributed to the improvement in labeling capability induced by GCN and BR.

(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Figure 3: Visual comparison of saliency maps with state-of-the-art methods, including our approach. (a) Input image (b) Ground truth (c) CARNet† (d) DCL (e) DHS (f) DSR (g) DRFI (h) HS (i) HDCT (j) MC. Our approach consistently produces saliency maps closest to the ground truth.

*	HKU-IS [22]		ECSSD [41]		PASCAL-S [18]		DUT-OMRON [32]	
	F-measure	AUC	F-measure	AUC	F-measure	AUC	F-measure	AUC
G-FRNet [8]	0.9085	0.9635	0.9080	0.9560	0.8310	0.9116	0.7840	0.9363
CARNet	0.9109	0.9657	0.9095	0.9567	0.8320	0.9142	0.7870	0.9405
CARNet†	**0.9115**	**0.9660**	**0.9250**	**0.9598**	**0.8341**	**0.9152**	**0.7895**	**0.9407**

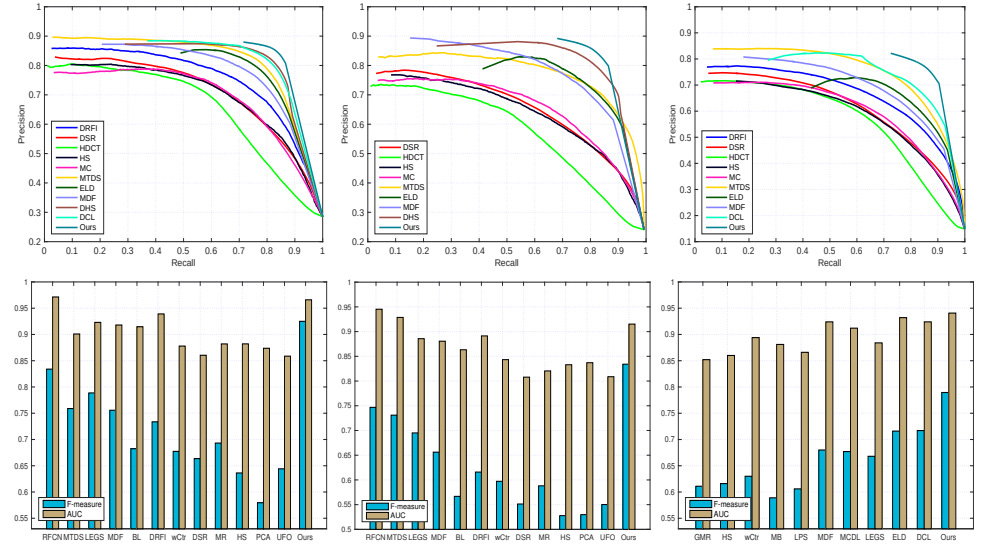
Table 2: Comparison of different variants of our proposed approach. Our final model CARNet† achieves the best performance when compared to other variants of our model.

5 Conclusion

In this paper, we have introduced a novel end-to-end refinement based architecture combined with prior information for solving the problem of salient object detection. Initially, the network generates a coarse label map by detecting the salient objects from a global view, then progressively recovers image details by integrating local context during the refinement. The most important contribution of our work is the stage-wise saliency map refinement, which results in precise saliency map. Experimental results demonstrate that the proposed model achieves state-of-the-art performance on benchmark datasets in salient object detection.

Acknowledgments

This work was supported by NSERC. We thank NVIDIA for donating some of the GPUs used in this work.



(a) ECSSD (b) PASCAL-S (c) DUT-OMRON

Figure 4: Comparison with state-of-the-art salient object detection methods on 3 different datasets. For each dataset, the first row shows the precision-recall curves and second row shows the F-measure and AUC. Our proposed approach CARNet† consistently outperforms other methods across all the datasets. In particular, the PR-curves show that our approach achieves significantly higher precision with higher recall, which demonstrates that our method locates salient objects more accurately and precisely. PR curves of our method terminate earlier than the baselines due to very high contrast (confidence) expressed in our predictions, that always achieves recall higher than 0.5. Note that DHSNet [19] includes the test set of DUT-OMRON in its training data. Therefore, we do not include it in the comparison based on the DUT-OMRON dataset.

References

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv:1606.00915*, 2016.

[2] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *TPAMI*, 37(3), 2015.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[4] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip Torr. Deeply supervised salient object detection with short connections. In *CVPR*, 2017.

[5] Md Amirul Islam, Shujon Naha, Mrigank Rochan, Neil Bruce, and Yang Wang. Label refinement network for coarse-to-fine semantic segmentation. *arXiv:1703.00551v1*, 2017.

[6] Md Amirul Islam, Mrigank Rochan, Neil Bruce, and Yang Wang. Gated feedback refinement network for dense image labeling. In *CVPR*, 2017.

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

[8] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv:1408.5093*, 2014.

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

[10] Dominik A Klein and Simone Frintrop. Center-surround divergence of feature statistics for salient object detection. In *ICCV*, 2011.

[11] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhenyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *AISTATS*, 2015.

[12] Gayoung Lee, Yu-Wing Tai, and Junmo Kim. Deep saliency with encoded low level distance map and high level features. In *CVPR*, 2016.

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

[14] Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In *CVPR*, 2016.

[15] Xi Li, Yao Li, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Contextual hypergraph modeling for salient object detection. In *ICCV*, 2013.

[16] 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. *TIP*, 25(8):3919–3930, 2016.

[17] Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via dense and sparse reconstruction. In *ICCV*, 2013.

[18] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *CVPR*, 2014.

[19] Nian Liu and Junwei Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *CVPR*, 2016.

[20] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *TPAMI*, 33(2):353–367, 2011.

[21] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[22] Vida Movahedi and James H Elder. Design and perceptual validation of performance measures for salient object segmentation. In *CVPRW*, 2010.

[23] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters–improve semantic segmentation by global convolutional network. *arXiv preprintarXiv:1703.02719*, 2017.

[24] Pedro O Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. Learning to refine object segments. In *ECCV*, 2016.

[25] Keyang Shi, Keze Wang, Jiangbo Lu, and Liang Lin. Pisa: Pixelwise image saliency by aggregating complementary appearance contrast measures with spatial priors. In *CVPR*, 2013.

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

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

[28] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Xiang Ruan. Saliency detection with recurrent fully convolutional networks. In *ECCV*, 2016.

[29] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, 2015.

[30] Yulin Xie, Huchuan Lu, and Ming-Hsuan Yang. Bayesian saliency via low and mid level cues. *TIP*, 22(5):1689–1698, 2013.

[31] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *CVPR*, 2013.

[32] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013.

[33] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Saliency detection by multi-context deep learning. In *CVPR*, 2015.