Hierarchical Salient Object Detection Network with Dense Connections

Qing Zhang(✉), Jianchen Shi, Baochuan Zuo, Meng Dai, Tianzhen Dong, and Xiao Qi

Shanghai Institute of Technology, Shanghai 201418, China
zhangqing0329@gmail.com

Abstract. Fully convolutional neural networks (FCNs) have shown outstanding performance in many dense labeling tasks. FCN-like salient object detection models haven mostly developed lately. In the work, we propose a novel pixel-wise salient object detection network based on FCN by aggregating multi-level feature maps. Our model first makes a coarse prediction by automatically learning various saliency cues, including color and texture contrast, shapes and objectness. Then a densely connected feature extraction block is adopted to further extract rich features at each resolution. Moreover, skip-layer structure is introduced for providing a better feature representation and helping shallow side outputs locate salient objects. In addition, a weighted-fusion module is utilized to combine multi-level features. Finally, a fully connected CRF model can be optimally incorporated to improve spatial coherence and contour localization in the fused saliency map. The whole architecture works in a coarse to fine manner. Evaluations on five benchmark datasets and comparisons with 10 state-of-the-art algorithms demonstrate the robustness and efficiency of our proposed model.

Keywords: Salient object detection · Visual saliency detection · Deep learning · Feature extraction

1 Introduction

Salient object detection aims at modeling human visual attention mechanism to segment the most distinct regions or objects from the clutter backgrounds. It has received a great deal of attention in computer vision community because of its wide range of applications including video summarization [1], content-aware image cropping and resizing [3,4] and person re-identification [2].

Since the seminal approaches of Itti et al. [5] and Liu et al. [6] are reported, extensive visual saliency algorithms have been proposed to simulate human

This work is supported by Natural Science Foundation of Shanghai under Grant Nos. 19ZR1455300 and 19ZR1455200, Science and Technology Development Foundation of Shanghai Institute of Technology under Grant No. ZQ2018-23, and National Natural Science Foundation of China under Grant No. 61806126.

Y. Zhao et al. (Eds.): ICIG 2019, LNCS 11901, pp. 454–466, 2019.
https://doi.org/10.1007/978-3-030-34120-6_37

visual attention mechanism in images and videos. Traditional salient object detection methods [7–10] adopt heuristic priors and manually designed features which are usually considered as low-level information. These generic techniques are useful for keeping fine images structures. However, these models cannot generate satisfied predicted results and are less applicable to a wide range of problems in practice. For example, it is difficult to pop out the salient objects when the background and salient objects share similar attributes (See the first row of Fig. 1). Moreover, it might fail sometimes, when there are multiple salient objects (See the second row of Fig. 1).

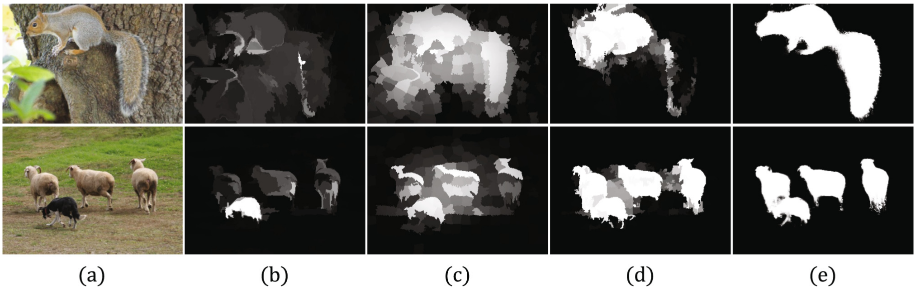


Fig. 1. Comparisons of results of different kinds of methods. For input images in (a), we show the salient object detection results of methods based on handcrafted features in (b) [10] and (c) [8], and salient object detection results of methods based on deep features in (d) [25] and (e) Ours.

In recent years, fully convolutional networks have shown powerful ability of feature representation and obtained impressive results in many dense labeling tasks including semantic segmentation [11,12], edge detection [14,15] and pose estimation [13]. Inspired by these achievements, researchers in the saliency detection community attempt to utilize its ability of adaptively extracting semantic features from raw images. These FCN-based models [16–18] have been successful in overcoming the disadvantages of handcrafted feature-based approaches and capturing high-level information about the objects and their clutter background, thus achieving better performance. However, although the saliency model using high-level information is superior, the low-level and mid-level features are also important in detecting salient objects. Therefore, it is a key and challenging issue to effectively and simultaneously aggregate multi-level saliency cues in a unified learning framework for capturing both the semantic objectness and detailed structure.

Motivated by these discussions, we propose a simple but effective salient object detection model for the pixel-wise saliency prediction task to simultaneously aggregate multi-level features to capture distinctive objectness and detailed information on complex images.

The main contributions are summarized as follows:

(1) A novel FCN-based saliency detection network model is proposed, which aggregates multi-level features as saliency cues. It performs image-to-image prediction and learns powerful and rich feature representations on complex images.
(2) We utilize the skip-layer scheme to guide low-level feature learning. With the help of deeper side information, shallower side outputs refine their predictions with more accurate location.
(3) The proposed model achieves state-of-the-art performance both quantitatively and qualitatively on DUT-OMRON [9], ECSSD [20], HKU [21], PASCAL-S [19] and SOD [34] benchmark datasets in terms of PR curves, F-measure, weighted F-measure and MAE scores.

2 Related Work

Generally, visual saliency detection approaches can be roughly classified into two categories: human fixation prediction and salient object detection. The former [5] is originally proposed to predict the fixation of eye movement, whereas the latter aims to detect and segment each entire salient object with explicit object boundaries from surroundings. Since this paper is focused on salient object detection based on deep learning, we will briefly review existing representative approaches for salient object detection.

2.1 Handcrafted Features Based Models

The majority of salient object detection approaches usually utilize handcrafted pixel/superpixel-level features, such as color, texture and orientation, by either local or global manner. The local based methods use rarity, contrast or distinctiveness of each pixel/region to capture the pixels/regions locally standing out from their surroundings, while the global based methods estimate the saliency of each pixel or region by using holistic priors of the entire image. Some researchers propose to build graphical models of superpixels to implicitly compute contrast [9,20]. They compute saliency by means of background, center, and compactness priors. However, traditional approaches, which mainly rely on handcrafted features, cannot describe semantic feature representation, therefore, they may fail to pop out salient objects in complex images.

2.2 Deep Neural Networks Based Models

Recently, deep learning based approaches, in particular the convolutional neural networks (CNNs), have been applied to detect salient objects and have improved the performance by a large margin. Wang et al. [23] propose one deep neural network to compute saliency score for each pixel in local context first, and then refine the saliency score for each object proposal over the global view with another network. Li et al. [21] predict saliency score of each superpixel by incorporating multi-scale features in a generic convolutional neural network. Zhao et al. [31] compute saliency by integrating global and local context into a deep learning based framework. Although these models achieve better results than traditional schemes, these models are very time-consuming due to the reason that they take segmented region as a basic unit to train a deep neural network for predicting saliency and the networks have to run many times for predicting saliency degree of all the superpixels in the image.

To remedy above problems, researchers prefer to adopt FCN-like model to detect saliency in a pixel-wise manner. Some researchers propose to use specific-level features for saliency prediction. For example, Lee et al. [25] propose to encode low-level distance map and high-level semantic features of deep CNNs. In [26], a network sharing features for segmentation and saliency tasks is proposed, and a graph Laplician regularized nonlinear regressor model is presented for refinement.

In contrary to these methods only use specific-level features, several works explore to integrate features from different side outputs and indicate that the features from all levels are potential saliency cues and are helpful for saliency prediction. The features from deep layers contain semantic information which is helpful for objectness, while the features from shallow layers contain rich detailed information which is helpful for explicit boundary in high-resolution prediction.

However, how to effectively and efficiently aggregate multi-level convolutional features remains challenging. To this end, several researchers make valuable attempts to solve this problem. Li et al. [27] combines a pixel-level fully convolutional stream and segmented-wise spatial pooling stream. The fully convolutional stream is a multi-scale fully convolutional network, which generates a saliency map with one eighth resolution of the raw input image by exploiting visual contrast across multiscale convolutional layers. Long et al. [11] introduce skip connections and adds high-level prediction layers to intermediate layers to generate pixel-wise prediction results at multiple resolutions. Liu et al. [16] design a two-stage deep network, in which a coarse global prediction is obtained by automatically learning various global structured saliency cues and another network is adopted to further refine the details of saliency maps via integrating local context information.

Though obvious achievement has been made by these deep learning based models in recent years, there is still a large room for improvement over the generic FCN-based models to uniformly highlight the entire salient objects and preserve the detailed boundaries against the cluttered background.

3 Proposed Model

Our proposed salient object detection model mainly consists of two stages: (1) a FCN-based deep network for multi-level features extraction and aggregation; and (2) a spatial coherence scheme for saliency refinement.

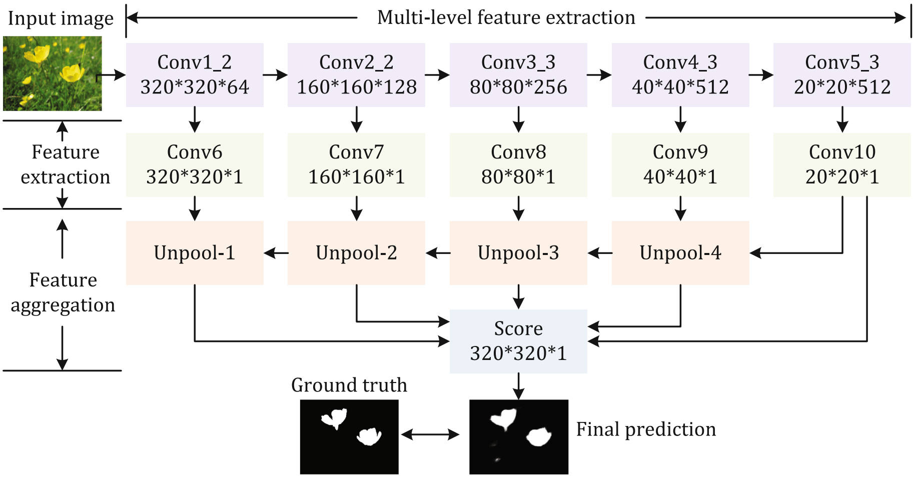


Fig. 2. The architecture of the proposed model. In the VGG-16 net, the names of the layers whose features are utilized are shown. The resolution of each step is also shown.

3.1 Network Architecture

To design a FCN-like network that is capable of accounting for both local and global context of an image and incorporating details from various resolutions, we develop a multi-scale deep convolutional neural network for learning discriminant saliency features (our mode is shown in Fig. 2). It consists of two components: feature extraction and aggregation.

Multi-level Feature Extraction. Our proposed model adopts VGG-16 net [28] (pre-trained over the ImageNet dataset for image classification) as our base network, and modifies it to meet our requirements. We retain its 13 convolutional layers, and remove the original 5th pooling layer and fully connected layers. Thus, the modified VGG-16 is composed of 5 groups of convolutional layers. For simplicity, we denote the third sub-layer in the fifth group of convolutional layer as $Conv5\_3$, and the other convolution layers in the VGG-16 is also denoted by this analogy. For an input image I with size $W \times H$, the modified VGGNet produces five feature maps f_i with decreasing spatial resolution by stride 2.

For each continuous feature f_i, $i \in \{5, 6, \ldots, 10\}$ extracted from VGG-16, we design a densely connected feature extraction block $Convi$. It utilizes a simple connectivity pattern: to preserve the feed-forward nature, each layer obtains additional inputs from all preceding layers and passes on its own feature maps to all subsequent layers, which is similar to DenseNet [24]. Figure 3 illustrates this layout schematically.

Features Aggregation. We obtain five feature maps with size different resolution from feature extraction blocks. The feature maps of deeper convolutional layers can accurately locate salient objects, while the feature maps generated by

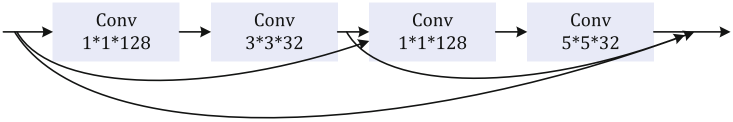


Fig. 3. Details of the feature extraction module.

shallower convolutional layers contain more details. To help the shallow side output contain more global properties, we refine these feature maps by skip-layer structure, namely, introducing the deeper side-output to its former shallower one. At each Unpool processing block, we combine features through summation. Moreover, we use a score module to integrate different maps and obtain a fused saliency map. To make the output maps of the features at different solutions have the same size for fusing, we use the deconvolutional layer for up-sampling. The strides of the last deconvolutional layers in the last four sides are respectively set to 2, 4, 8 and 16. And then, we combine features by concatenating them.

3.2 Spatial Coherence

To improve spatial coherence and achieve more accurate results, we adopt a pixel-wise saliency refinement model based on a fully connected conditional random field (CRF) [29] in the inference phase. This CRF model solves a binary pixel labeling problem, which is similar to our saliency prediction task, and employs the following energy function,

$$E(L) = -\sum_{i} logP(l_i) + \sum_{i,j} \theta_{ij}(l_i, l_j) \tag{1}$$

where L represents a binary label assignment for all pixels. $P(l_i)$ is the probability of pixel x_i with label l_i, which indicates the likelihood of pixel x_i being salient. Initially, $P(1) = S_i$ and $P(0) = 1 - S_i$, where S_i is the saliency score at pixel x_i from the fused saliency map S. $\theta_{i,j}(l_i, l_j)$ is a pairwise potential and defined as follows,

$$\theta_{ij} = \mu(l_i, l_j)[\omega_1 exp(-\frac{||p_i - p_j||^2}{2\sigma_\alpha^2}) - \frac{||I_i - I_j||^2}{2\sigma_\beta^2} + \omega_2 exp(-\frac{||p_i - p_j||^2}{2\sigma_\gamma^2})] \tag{2}$$

where $\mu(l_i, l_j) = 1$ if $l_i \neq l_j$, and zero otherwise. θ_{ij} involves two kernels. The first kernel depends on pixel positions p and pixel intensities I. This kernel makes nearby pixels having similar colors take similar saliency scores. Three parameters determine the degree of influence by color similarity and spatial relation, respectively. The second kernel is to remove small isolated regions. The parameters of ω_1, ω_2, σ_α^2, σ_β^2, σ_γ^2 are set to 3.0, 3.0, 60.0, 8.0 and 5.0 respectively in our experiments.

4 Experiments

4.1 Implementation Details

Our network is based on the publicly available Caffe library, an open source framework for CNNs training and testing. As mentioned above, we choose VGG-16 as our pre-trained model and fine-tune it for pixel-wise saliency prediction. We utilize the same training and validation sets as in [8]. The learning rate is set to 1e−9, the momentum parameter is 0.9, the weighted decay is set to 0.0005. The fusion weight in the feature integration module are all initialized with 0.2 in the training phase.

4.2 Datasets

We conduct evaluations on five widely used salient object benchmark datasets. DUT-OMRON is manually selected from more than 140,000 natural images, each of which has one or more salient objects and relatively complex backgrounds. As an extension of the Complex Scene Saliency Dataset (CSSD), ECSSD is obtained by aggregating the images from two publicly available datasets and the Internet. HKU contains 4447 images, most of which have low contrast and multiple salient objects. PASCAL-S is generated from the PASCAL VOC dataset with 20 object categories and complex scenes. SOD is more challenging with multiple salient object and background clutters in images.

4.3 Evaluation Metrics

We adopt the precision-recall (PR) curve to evaluate our proposed model. The precision and recall are computed by binarizing the saliency map with 256 thresholds, ranging from 0 to 255, and comparing the binary map with the ground truth. The PR curves demonstrate the mean precision and recall of saliency maps at different thresholds. We also use F-measure (F_β) and weighted F-measure ($\omega\mathrm{F}_\beta$) scores to comprehensively consider precision and recall. F_β is given by:

$$F_\beta = \frac{(1+\beta^2)\cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \tag{3}$$

where β is a balance parameter to weight the precision and recall, and β^2 is set to 0.3. Similar to F_β, $\omega\mathrm{F}_\beta$ is computed with a weighted harmonic mean of $Precision^w$ and $Recall^w$: $F_\beta^w = \frac{(1+\beta^2)\cdot Precision^w \cdot Recall^w}{\beta^2 \cdot Precision^w + Recall^w}$.

Beside, we use the mean absolute error (MAE) to evaluate the average pixel-wise error between the saliency map and ground truth. It is defined as $MAE = \frac{1}{h\cdot w}\sum_{i=1}^{h}\sum_{j=1}^{w}|S_{ij} - G_{ij}|$ where S denotes the saliency map, G denotes the ground truth, and h and w denote the height and width of the image.

4.4 Performance Comparison with State of the Art

We compare our proposed approach with 10 state-of-the-art methods, including UCF [33], MTDS [26], LEGS [23], MDF [21], KSR [30], DRFI [8], SMD [10], ELD [25], MC [31], and ELE [32]. We use either the implementations or the saliency maps provided by the authors for fair comparison. Note that MC, UCF, ELD, MTDS, LEGS, MDF, KSR are deep learning based models.

Table 1. F_β and ωF_β scores of saliency maps produced by different approaches on DUT-OMRON, ECSSD, HKU, PASCAL-S, and SOD datasets (The top models are highlighted in bold. '-' denotes the saliency maps are not available).

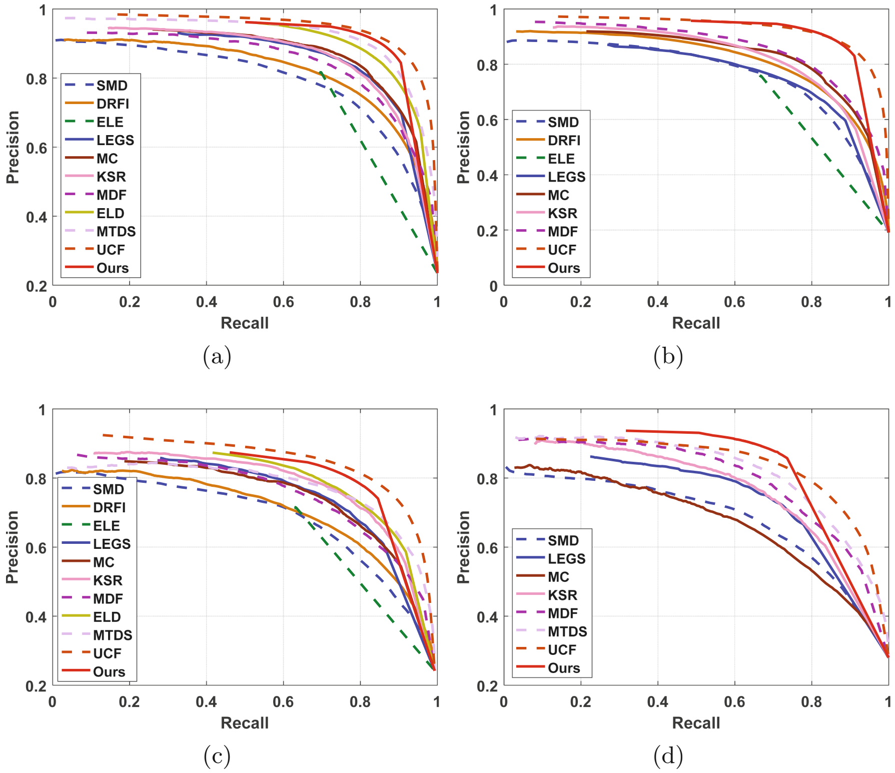
Approach	DUT-OMRON		ECSSD		HKU		PASCAL-s		SOD	
	F_β	ωF_β	F_β	ωF_β	F_β	ωF_β	F_β	ωF_β	F_β	ωF_β
SMD	0.537	0.398	0.712	0.532	0.691	0.499	0.622	0.462	0.605	0.474
DRFI	0.555	0.374	0.732	0.567	0.722	0.502	0.613	0.446	-	-
ELE	0.575	0.525	0.755	0.720	0.699	0.655	0.652	0.604	-	-
LEGS	-	-	0.783	0.723	0.709	0.616	0.688	0.610	0.686	0.612
MC	-	-	0.797	0.750	0.759	0.700	0.692	0.628	0.589	0.391
KSR	0.591	0.493	0.782	0.675	0.747	0.638	0.703	0.610	0.668	0.579
MDF	0.596	0.499	0.749	0.643	0.764	0.641	0.648	0.557	0.697	0.601
ELD	0.614	0.564	0.817	0.773	-	-	0.721	0.659	-	-
MTDS	0.603	0.463	0.826	0.693	-	-	0.658	0.521	0.698	0.568
UCF	0.621	0.537	0.844	0.788	0.823	0.754	0.733	0.669	0.738	0.684
Ours	**0.660**	**0.615**	**0.862**	**0.851**	**0.868**	**0.845**	**0.747**	**0.719**	**0.759**	**0.759**

For quantitative evaluation, we show comparison results with PR curves and MAE scores in Figs. 4 and 5. And the comparisons of F_β and ωF_β are displayed in Table 1. We do not show the comparison of PR curves on DUT-OMRON due to the limited space. In terms of F_β, ωF_β and MAE scores, we can see that our model outperforms all other methods, especially on complex datasets. For the PR curves, our model also achieves a good performance on four datasets and is a little worse than UCF on ECSSD and PASCAL-S.

We show visual comparison in Fig. 6. We can see that our model not only detects and localizes salient objects accurately, but also preserves object details subtly. It can handle various complex situations well, including salient objects being small (row fourth and fifth), clutter backgrounds and salient objects (row first and sixth), backgrounds and salient objects sharing similar appearance (row second, third and fifth).

4.5 Evaluation on CRF Scheme

A fully connected CRF scheme is incorporated to further uniformly highlight the interior regions of salient object and preserve explicit contour in the saliency map



(a)

(b)

(c)

(d)

Fig. 4. PR curves of saliency maps produced by different approaches on four datasets. (a) ECSSD, (b) HKU, (c) PASCAL-S and (d) SOD.

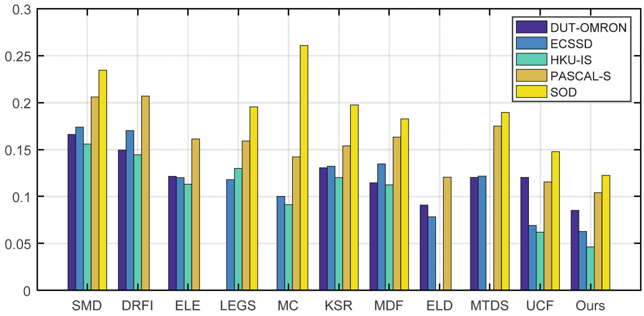


Fig. 5. MAE scores of the saliency maps produced by different models on five datasets. Lower is better.

from our proposed multi-scale FCN-like network. To validate its effectiveness, we have also evaluated the performance of our final saliency approach with and without (w/o) CRF scheme on five benchmark datasets in terms of F_β, ωF_β, and

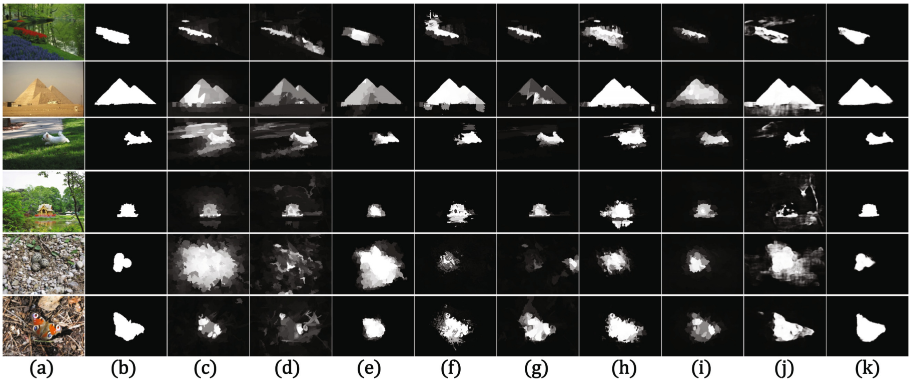


Fig. 6. Visual comparison results based on different models. (a) Input, (b) ground truth, (c) SMD, (d) DRFI, (e) LEGS, (f) MC, (g) MDF, (h) ELD, (i) MTDS, (j) UCF, and (k) Ours.

MAE scores. The results are displayed in Table 2, which shows that the CRF scheme improves the accuracy of our proposed model.

Table 2. Comparisons of our approach with and without(w/o) CRF scheme in terms of F_β, ωF_β, and MAE.

Datasets	Method	F_β	ωF_β	MAE
DUT-OMRON	Ours with CRF	0.6600	0.6152	0.0852
	Ours w/o CRF	0.6265	0.5753	0.0932
ECSSD	Ours with CRF	0.8621	0.8505	0.0627
	Ours w/o CRF	0.8299	0.8019	0.0730
HKU	Ours with CRF	0.8681	0.8454	0.0463
	Ours w/o CRF	0.8260	0.7897	0.0569
PASCAL-S	Ours with CRF	0.7465	0.7187	0.1041
	Ours w/o CRF	0.7180	0.6816	0.1127
SOD	Ours with CRF	0.7594	0.7589	0.1225
	Ours w/o CRF	0.7503	0.7303	0.1284

5 Conclusion

In this paper, we propose a simple but effective approach for pixel-wise salient object detection based on a fully convolutional network, which extracts multi-level features and utilizes the preceding information through a densely connected module. Moreover, the features from deeper layers are connected to the shallower ones by skip-layer structure for guiding the learning of shallower layers. Besides, a fusion layer is adopted to combine these rich features to generate a saliency map. In order to obtain more fine-gained saliency detection results, we introduce a saliency refinement scheme based on a fully connected CRF to further improve saliency performance. Experimental results demonstrate that our proposed approach achieves encouraging performance against 10 state-of-the-art methods on five benchmark datasets.

References

1. Simakov, D., Caspi, Y., Shechtman, E., Irani, M.: Summarizing visual data using bidirectional similarity. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
2. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3586–3593 (2013)
3. Wang, W., Shen, J., Yu, Y., Ma, K.: Stereoscopic thumbnail creation via efficient stereo saliency detection. IEEE Trans. Vis. Comput. Graph. **23**(8), 2014–2027 (2017)
4. Wang, W., Shen, J., Ling, H.: A deep network solution for attention and aesthetics aware photo cropping. IEEE Trans. Pattern Anal. Mach. Intell. **41**, 1 (2018)
5. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998)
6. Liu, T., Sun, J., Zheng, N., Tang, X., Shum, H.: Learning to detect a salient object. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
7. Perazzi, Y., Krahenbuhl, P., Hornung, H.: Saliency filters: contrast based filtering for salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 733–740 (2012)
8. Wang, J., Jiang, H., Yuan, Z., Cheng, M., Hu, X., Zheng, N.: Salient object detection: a discriminative regional feature integration approach. Int. J. Comput. Vis. **123**(2), 251–268 (2017)
9. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.: Saliency detection via graph-based manifold ranking. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3166–3173 (2013)
10. Peng, H., Li, B., Ling, H., Hu, W., Xiong, W., Maybank, S.: Salient object detection via structured matrix decomposition. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 818–832 (2017)
11. Long, J., Shellhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
12. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: International Conference on Computer Vision, pp. 1520–1528 (2015)
13. Yang, W., Ouyang, W., Li, H., Wang, X.: End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3073–3082 (2016)
14. Xie, S., Tu, Z.: Holistically-nested edge detection. In: International Conference on Computer Vision, pp. 1395–1403 (2015)
15. Liu, Y., et al.: Richer convolutional features for edge detection. IEEE Trans. Pattern Anal. Mach. Intell. (2019). https://doi.org/10.1109/TPAMI.2018.2878849

16. Liu, N., Han, J.: DHSNet: deep hierarchical saliency network for salient object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 678–686 (2016)
17. Zhang, P., Wang, D., Lu, H., Wang, H., Ruan, X.: Amulet: aggregating multi-level convolutional features for salient object detection. In: International Conference on Computer Vision, pp. 202–211 (2017)
18. Hou, Q., Cheng, M., Hu, X., Borji, A., Tu, Z., Torr, P.: Deeply supervised salient object detection with short connections. IEEE Trans. Pattern Anal. Mach. Intell. **41**(4), 815–828 (2019)
19. Li, Y., Hou, X., Koch, C., Rehg, J., Yuille, A.: The secrets of salient object segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 280–287 (2014)
20. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1155–1162 (2013)
21. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5455–5463 (2015)
22. Zhu, W., Liang, S., Wei, Y., Sun, J.: Saliency optimization from robust background detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2814–2821 (2014)
23. Wang, L., Lu, H., Yang, M.: Deep networks for saliency detection via local estimation and global search. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3183–3192 (2015)
24. Huang, G., Liu, Z., Maaten, L., Weinberger, K.: Densely connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2261–2269 (2017)
25. Lee, G., Tai, Y.W., Kim, J.: Deep saliency with encoded low level distance map and high level features. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 660–668 (2016)
26. Li, X., et al.: DeepSaliency: multi-task deep neural network mode for salient object detection. IEEE Trans. Image Process. **25**(8), 3919–3930 (2016)
27. Li, G., Yu, Y.: Deep contrast learning for salient object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 478–487 (2016)
28. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
29. Krahenbuhl, P., Koltun, V.: Efficient inference in fully connected CRFs with Gaussian edge potentials. In: Neural Information Processing Systems, pp. 109–117 (2011)
30. Wang, T., Zhang, L., Lu, H., Sun, C., Qi, J.: Kernelized subspace ranking for saliency detection. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 450–466. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_27
31. Zhao, R., Ouyang, W., Li, H., Wang, X.: Saliency detection by multi-context deep learning. In: IEEE Conference Computer Vision and Pattern Recognition, pp. 1265–1274 (2015)
32. Xia, C., Li, J., Chen, X., Zheng, A., Zhang, Y.: What is and what is not a salient object? Learning salient object detector by ensembling linear exemplar regressors. In: International Conference on Computer Vision and Pattern Recognition, pp. 4399–4407 (2017)

33. Zhang, P., Wang, D., Lu, H., Wang, H., Yin, B.: Learning uncertain convolutional features for accurate saliency detection. In: International Conference on Computer Vision, pp. 212–221 (2017)
34. Movahedi, V., Elder, J: Design and perceptual validation of performance measures for salient object segmentation. In: International Conference on Computer Vision and Pattern Recognition, pp. 49–56 (2010)