Salient Object Detection With CNNs and Multi-scale CRFs

Yingyue Xu, Xiaopeng Hong, and Guoying Zhao

Center for Machine Vision and Signal Analysis
University of Oulu
{yingyue.xu, xiaopeng.hong, guoying.zhao}@oulu.fi

Abstract. Recent CNNs based salient object detection approaches tend to embed a fully connected Conditional Random Field (CRF) layer to refine the saliency maps from CNNs for post processing. Due to the significant performance enhancement by the CRF layer, in this paper, we propose a more flexible CRF refinement framework by embedding the CRF inference to multiple levels of side outputs from CNNs for multi-scale saliency refinement. A fully convolutional neural networks based on the simple yet effective encoder-decoder architecture with only three scales of side output maps is pre-trained. Then, the CRF layers are embedded to each scale of the side output respectively to complement the defects of each side output maps. Finally, the refined side output maps are fused and refined by another CRF inference for the final saliency map. The proposed multi-scale CRFs model (MCRF) is trained with low computational costs and shows competitive performance over four datasets in comparison with the existing state-of-the-art saliency models.

Keywords: Saliency detection · CRF · Multi-scale FCNs.

1 Introduction

Salient Object Detection (SOD) refers to the perceptual selective process that highlights the most distinct regions on the scenes by the human vision system. In practice, object-level saliency detection can be broadly applied as a pre-processing technique for various computer vision tasks, such as image and video segmentation [27], video compression [6], image cropping [28], video summarizing [23], image fusion [7], *etc.*

Due to the prevalence of the convolutional neural networks (CNNs), the performance of salient object detection has been largely improved [19, 33, 16, 5, 18, 9, 17, 32, 24, 13]. Further, the integration of fully convolutional neural networks (FCNs) facilitates salient object detection tasks to an end-to-end phase [18, 34, 12, 30, 42]. However, current CNNs based SOD approaches still face ~~with~~ significant challenges due to its network structures. Firstly, as salient object detection is a pixel-level labelling task, the outputs from the convolutional layers with large receptive fields can be rather rough after being restructured back to pixel-level labelled maps [21]. Hence, the output saliency maps from CNNs may result in blob-like defects. Secondly, salient object detection emphasizes on recognizing the salient regions in object-level. As the CNNs lacks smoothness constraints for label agreement, the output saliency maps may consist poor object delineation and spurious regions [25]. In summary, the output maps from CNNs are relatively coarse and further refinement is needed to improve object boundary division and saliency density smoothness.

One prevalent approach adopted by recent state-of-the-art saliency models is to introduce a Conditional Random Field (CRF) layer, *i.e.* Dense-CRF [11], fully connected to the FCNs [17, 14, 9] for coarse saliency map refinement. The fully connected CRF layer does not participate in finetuning the front-end FCNs. Instead, it acts as a post-processing layer to reconcile the spatial and appearance coherence of the coarse saliency maps through cross validations. On one hand, the CRF layer efficiently enhances the accuracy of the saliency maps in practice; on the other hand, it makes the training stage of the front-end deep neural networks compact and efficient.

However, existing saliency models only connect one CRF layer to the end of the pre-trained deep neural networks for refinement. In this paper, we extend the CRF layer as a more flexible integration to any of the side output layers of the FCNs to enhance the quality of the intermediate outputs, and thus to further improve the performance of the whole networks. Therefore, we propose the multi-scale CRFs model (MCRF) based on multi-scale side outputs from FCNs for salient object detection. Specifically, a fully convolutional neural network based on the encoder-decoder architecture with three scales of side output maps is trained with pixel-wise labels. Then, a CRF layer is connected to each side output layer to refine the delineation and smoothness of the side output maps. Finally, the refined side output maps are fused and then refined by another CRF layer for the final saliency map. The contributions of the paper are two folds:

- The proposed MCRF model integrates multiple CRF layers to refine the multi-scale side output maps from FCNs, and thus to complement the defects of each side outputs for a unified and refined saliency map. The multi-scale CRFs refinement structure largely improves the refinement effectiveness than integrating one CRF layer at the end of the network.
- The multi-scale CRFs refinement structure results in highly competitive performance based on the simple encoder-decoder networks with only three scales of side outputs. Hence, the multi-scale CRFs structure is able to avoid the over-fitting issues due to complex deep network architectures with limited training samples.

The rest of the paper is organized as follows. Section 2 summarizes the related works. Section 3 introduces the framework of the proposed multi-scale CRFs saliency model. Section 4 presents the implementation details and the experimental results and Section 5 concludes the work.

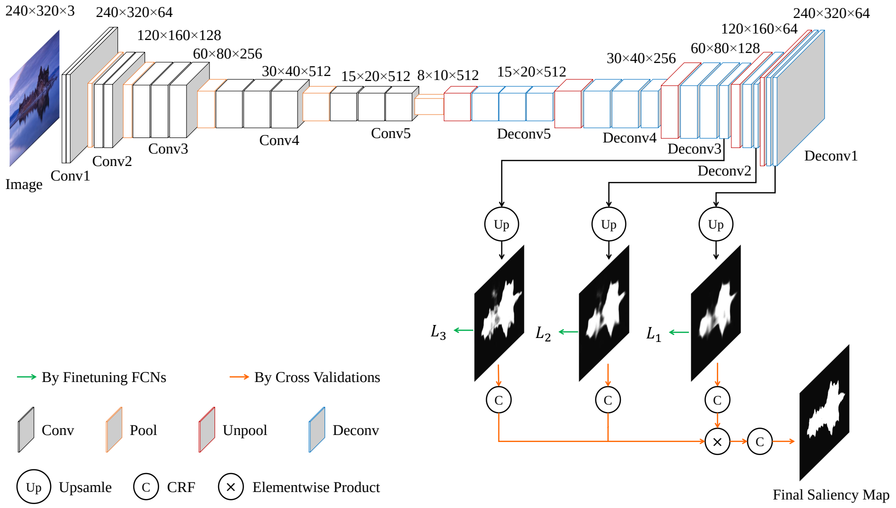

Fig. 1. Framework of the proposed multi-scale CRFs model. Three scales of side output maps are selected from the encoder-decoder networks. The encoder network is based on the VGG-16 net [31]. Then, the decoder network is connected to the "pool5" layer, which gradually unpools the features from the corresponding pooling layers. The decoder convolutional layers are all followed by a BN layer and a ReLU layer. To upsample the three scales of side outputs from "Deconv1", "Deconv2" and "Deconv3", a convolutional layer with 1×1 kernel size is used to compute the one channel feature map and a deconvolutional layer followed with a crop layer is connected to upample the feature maps to the image size respectively. To finetune the front-end encoder-decoder networks, each side output map is connected with a side loss (L_1, L_2, L_3) for optimization. Then, one CRF layer is connected to each side output map for multi-scale refinement and the refined side output maps are fused by element-wise production. Finally, another CRF layer is connected to refine the fused map for the final saliency map. The CRF layers are tuned by cross validations and all the CRF layers share the same parameter settings.

2 Related Works

This section presents a brief review of representative network architectures of FCNs and previous deep saliency models that adopt fully connected conditional random field (CRF) for saliency refinement.

2.1 Saliency Detection via FCNs

Previous works [41] suggest that the convolutional layers of the CNNs can describe the high-level semantic features at different scales and maintain their spatial information. To put the merits of convolutional layers into full use, Long *et al.* [21] propose the fully convolutional networks (FCNs) for semantic segmentation. The FCNs addresses the advantages of the convolutional layers to get rid of the large parameter costs from the fully connected layers. Moreover, as the convolutional layers keep the spatial information on the output maps, the side outputs from different levels are able to produce multi-scale feature maps for recognition tasks [2, 22].

A variety of network architectures are proposed to compute multi-scale feature maps of FCNs. The encoder-decoder network [40] proposes a simple yet efficient fully convolution and unpooling structure for object contour detection. The seminal FCNs [21] constructs skip connections to generate end-to-end prediction at multiple scales. Similarly, the Hypercolumns FCNs [8] and the U-Net [29] apply multiple skip-connections through concatenations to capture the features from multiple scales for precise localization. The holistically-nested edge detector (HED) model [35] employs skip-layer connections to construct even deeper supervised structures, and fuses side outputs at various scales to resolve the ambiguity in edge and object boundary detection. Further, the DSS model [9] introduces short connections to the skip layers to construct an enhanced HED structure that combines both deeper and shallower side outputs for multi-scale contexts. Apparently, deeper network architectures are able to learn richer semantic features for more accurate predictions. However, complex network architectures may lead to the time consuming training process and may face with over-fitting problems. Thus, constructing a compact yet efficient FCNs structure for targeted tasks is crucial in balancing the accuracy and efficiency.

2.2 CRFs for Saliency Refinement

Prior to the pervasive applications of the CNNs, most of the best performed traditional saliency methods firstly compute a coarse saliency map and then refine it by handcraft features from the input image. Such refinement is based on some common context-aware assumptions and theories from graphical models. As the conditional random fields (CRFs) is a flexible framework for incorporating various features and is capable to accommodate inference functions for graphical models, it has been frequently adopted for labeling refinement tasks. For instance, Qiu *et al.* [26] take advantages of handcraft image features and spatially weighted distance to infer a CRF model to refine coarse saliency maps.

Deeplab [3] firstly implements the dense CRFs framework to deep neural networks to refine the semantic segmentation results based on unary and pairwise potentials proposed by [11]. The proposed dense CRFs is fully connected to CNNs as a post-processing step for end-to-end refinement. Later, several works [44, 37, 36] unroll the CRF inference by [11] to an end-to-end trainable feed-forward networks.

For efficient computation, existing saliency models tend to integrate the fully connected CRFs on top of the deep neural networks for end-to-end post processing. MDF saliency model [15] involves the CRFs from Deeplab [3] to integrate multiple output saliency maps from CNNs with inputs of different contexts. Later, the DCL model [16] incorporates the dense CRFs [3] to improve spatial coherence and contour localization for the fused result from two streams of

CNNs. The MSRNet [14] model and DSS [9] model both integrates the fully connected dense CRFs [11] to refine the fused output maps from the CNNs. In this work, a more flexible and efficient incorporation of dense CRFs will be explored on top of the pre-trained CNNs.

3 Multi-scale CRFs Model

The Multi-scale CRFs Model is based on a simple yet effective encoder-decoder architecture. Firstly, multi-scales of side output feature maps are computed from the pre-trained encoder-decoder networks. Then, each side output maps are refined by a fully connected CRF layer to enhance the delineation and smoothness. Finally, the enhance feature maps are fused and refined by another CRF layer for the final saliency map.

3.1 The Encoder-decoder Networks

Given the input image $I = \{I_i, i = 1, \cdots, |I|\}$ with three-dimensional size of $H \times W \times 3$, and the ground truth $G = \{G_i, i = 1, \cdots, |G|\}$, $G_i \in \{0, 1\}$ with the size of $H \times W \times 1$, the encoder-decoder networks $\mathcal{F}$ is adopted to produce $M = \{m = 1, \cdots, M\}$ scales of side output feature maps, denoted as s_m respectively as follow:

$$s_m = \mathcal{F}(W, w_m), \tag{1}$$

where W denotes the generic weights of the encoder-decoder networks and w_m denotes the scale specific weights. In the training phase, the cross-entropy loss is utilized as the side objective function $L_m(W, w_m)$ to train the network weights:

$$\begin{aligned} & L_m(W, w_m) \\ & = -\sum_{\mathcal{S}_m^j \in \mathcal{S}_m} [\mathcal{S}_m^j \log P(\mathcal{S}_m^j = 1|\mathcal{I}; W, w_m) + (1 - \mathcal{S}_m^j) \log P(\mathcal{S}_m^j = 0|\mathcal{I}; W, w_m)] \end{aligned} \tag{2}$$

where $\mathcal{I} = \{\mathcal{I}^j, j = 1, \cdots, |\mathcal{I}|\}$ denotes all the pixels in the training image set and $\mathcal{S}_m = \{\mathcal{S}_m^j, j = 1, \cdots, |\mathcal{S}_m|\}$ denotes all the saliency values from the side output layer at the m-th scale of the encoder-decoder networks. $P(\mathcal{S}_m^j = 1|\mathcal{I}; W, w_m)$ represents the probability of the activation value at location j at the m-th scale side output map.

3.2 Multi-scale CRFs Refinement

Through the encoder-decoder networks, M scales of side output maps are computed to primarily locate the salient objects. In order to further improve the prediction accuracy, a fully connected CRF[11] layer is integrated to each side output layer for refinement as follow:

$$\hat{s}_m = \mathcal{C}_m(s_m, I, \Theta_m), \tag{3}$$

where $\mathcal{C}_m(\cdot)$ refers to the CRF layer at the m-th scale, Θ_m refers to all the parameters for the m-th CRF layer, and $\hat{s}_m$ represents the refined side output map at the m-th scale.

To each side output map $\hat{s}_m$, the energy function of the CRF is

$$E(G) = \sum_i \phi_u(s_m^i) + \sum_{i<k} \phi_p(s_m^i, s_m^k). \tag{4}$$

$\phi_u(s_m^i)$ refers to the unary term, where the side output maps are directly regarded as the input. $\phi_p(s_m^i, s_m^k)$ is the pairwise term, which accounts for the coherence of the saliency information and image features between the current pixel and its neighbors. Thus, the pairwise term is defined as:

$$\phi_p(s_m^i, s_m^k) = \mu(s_m^i, s_m^k)[\nu_1 \exp(-\frac{\|p_i - p_k\|^2}{2\sigma_\alpha^2} - \frac{\|I_i - I_k\|^2}{2\sigma_\beta^2}) + \nu_2 \exp(-\frac{\|p_i - p_k\|^2}{2\sigma_\gamma^2})], \tag{5}$$

where $\mu(s_m^i, s_m^k) = 1$ if $s_m^i = s_m^k$ and otherwise 0. I_i represents the RGB image features of the i-th pixel, while p_i is the pixel position. The Gaussian kernel $\exp(-\frac{\|p_i - p_k\|^2}{2\sigma_\alpha^2} - \frac{\|I_i - I_k\|^2}{2\sigma_\beta^2})$ measures the appearance coherence which refines the nearby pixels with similar features with similar saliency scores, while the Gaussian kernel $\exp(-\frac{\|p_i - p_k\|^2}{2\sigma_\gamma^2})$ measures the spatial coherence which reconciles close pixels with similar saliency scores. Parameters ν_1 and ν_2 control the contributions of each Gaussian kernel respectively.

The energy minimization is based on the mean field approximation to the CRF distribution proposed by [11], and high-dimensional filtering can be utilized to speed up the computation.

Then, the refined saliency maps from each scale of the CRF layer are fused by element-wise production:

$$\tilde{s} = \prod_{m=1}^{M} \hat{s}_m. \tag{6}$$

Finally, another CRF layer is connected to further refine the fused map as the final saliency map:

$$\bar{s}_{final} = \mathcal{C}_{final}(\tilde{s}, I, \Theta_{fuse}) \tag{7}$$

4 Experiment

4.1 Implementation

In this work, the fully convolutional encoder-decoder networks is adopted to obtain the multi-scale side output maps. The network architecture is demonstrated

Dataset	Metric	MDF$^+$	RFCN	DHS	Amulet	UCF	DCL$^+$	MSR$^+$	DSS$^+$	RA	MCRF
MSRA-B	$\mathtt{F}_\beta$	0.862	-	-	-	-	0.898	0.880	**0.917**	0.901	**0.917**
	MAE	0.066	-	-	-	-	0.047	0.053	**0.034**	0.050	0.038
DUT-	$\mathtt{F}_\beta$	0.677	0.693	-	0.668	0.630	0.716	0.695	**0.757**	0.719	0.726
OMRON	MAE	0.092	0.095	-	0.098	0.120	0.080	0.078	**0.063**	0.076	0.074
HKU-IS	$\mathtt{F}_\beta$	0.809	0.868	0.863	0.849	0.816	0.851	0.864	**0.904**	0.879	0.894
	MAE	0.076	0.073	0.052	0.050	0.061	0.049	**0.040**	**0.040**	0.057	**0.040**
ECSSD	$\mathtt{F}_\beta$	0.818	0.871	0.875	0.865	0.835	0.874	0.838	**0.899**	0.876	0.893
	MAE	0.106	0.091	0.059	0.061	0.071	0.068	0.062	**0.055**	0.080	0.058

Table 1. Evaluation results over four datasets, with models including MDF [15], RFCN [34], DHS [18], Amulet [42], UCF [43], DCL [16], MSR [14], DSS [9], RA [4] and the proposed MCRF model. "+" marks the models utilizing dense CRF [11] for post-processing. "-" means that the corresponding dataset is used as the training data. The evaluation on MSRA-B is performed on the testing set. The best performances are in **bold** while the second best results are underlined.

in Figure 1 with detailed layer descriptions. The encoder network is based on the VGG-16 net [31]. The decoder network firstly unpools the features from the corresponding maxpooling layers and properly upsample and crop the side output maps to the image size. All the decoder convolutional layers are followed by batch normalization and ReLU activation functions. We also add a dropout layer after each ReLU layer in the decoder networks.

The hyper-parameters for the finetuning the encoder-decoder networks are set as: a fixed learning rate (1e-8), weight decay (0.0005), momentum (0.9), loss weight for each side output (1). The batch size is set as 12, and 100 epochs are performed for tuning the encoder decoder network. The sigmoid cross entropy loss layers are used for model optimization.

The fully connected dense CRF layers share the same parameter settings and are tuned via cross validations on the validation set, and ν_1, ν_2, σ_α, σ_β, and σ_γ are set to 3.0, 3.0, 60.0, 5.0, and 3.0, respectively. Only 3 iterations of the meanfield approximation are set to each CRF layer.

All the implementation is based on the public Caffe library [10]. The CRF is based on the the PyDenseCRF implementation [11]. The GPU for training acceleration is the Nvidia Tesla P100 with 16 GB memory. Totally 100 epochs are performed to train the encoder-decoder networks, which takes about 16 hours. In the testing phase, it takes averagely 1.68s to compute the final saliency maps.

4.2 Datasets

We follow the training protocol as in [9, 16] by using the MSRA-B dataset [20] as the training data for fair comparisons. The MSRA-B dataset consists of 2,500 training images, 500 validation images and 2000 testing images. The images are resized to 240×320 as the input to the data layer. Horizontal flipping is used for data augmentation such that the number of training samples is twice as large as the original number.

Maps	s_1	s_2	s_3	s_{123}+$\texttt{CRF}^1$	s_{123}+$\texttt{CRF}^2$	MCRF
$\texttt{F}_\beta$	0.856	0.849	0.839	0.867	0.868	0.893

Table 2. Comparisons of Mean F-measure by implementing multi-scale CRFs versus implementing single-scale CRF respectively. "s_1, s_2, s_3" refer to the three scales of side output maps from the encoder-decoder networks respectively. "s_{123}+$\texttt{CRF}^1$" fuses the maps "s_1, s_2, s_3" by elementwise production and then connect a single CRF layer with 3 meanfield iterations to compute the saliency maps. "s_{123}+$\texttt{CRF}^2$" also fuses the side output maps and connect a single CRF layer with 10 meanfield iterations. Note that the parameter settings of CRF layer for "s_{123}+$\texttt{CRF}^2$" are the same as DSS model. The evaluations are performed on ECSSD dataset.

The proposed model is evaluated over four datasets, including: MSRA-B [20], ECSSD [38], DUT-OMRON [39], and HKU-IS [17]. MSRA-B is the training dataset. ECSSD contains a pool of 1000 images with even more complex salient objects on the scenes. DUT-OMRON dataset contains a large number of 5168 more difficult and challenging images. HUK-IS consists of 4447 challenging images and pixel-wise saliency annotation.

4.3 Evaluation Metrics

We employ two types of evaluation metrics to evaluate the performance of the saliency maps: mean F-measure and mean absolute error (MAE). When a given saliency map is slidingly thresholded from 0 to 255, a precision-recall (PR) curve can be computed based on the ground truth. F-measure is computed to count for the saliency maps with both high precision and recall:

$$F = \frac{\left(1+\beta^2\right) \cdot precision \cdot recall}{\beta^2 \cdot precision + recall}, \tag{8}$$

where $\beta^2 = 0.3$ [1] to emphasize the precision. In this paper, the mean F-measure is chosen for evaluation and the saliency maps are thresholded by twice of the mean saliency values.

MAE measures the overall pixel-wise difference between the saliency map sal and the ground truth gt as follow:

$$MAE = \frac{1}{H}\sum_{h=1}^{H} |sal(h) - gt(h)|, \tag{9}$$

where H is the number of pixels on the map.

4.4 Experimental Results

We compare the proposed MCRF model with nine state-of-the-art deep saliency models including MDF [15], RFCN [34], DHS [18], Amulet [42], UCF [43], DCL [16], MSR [14], DSS [9], and RA [4]. All the models are CNN-based approaches. All the implementations are based on public codes and suggested settings by the corresponding authors.

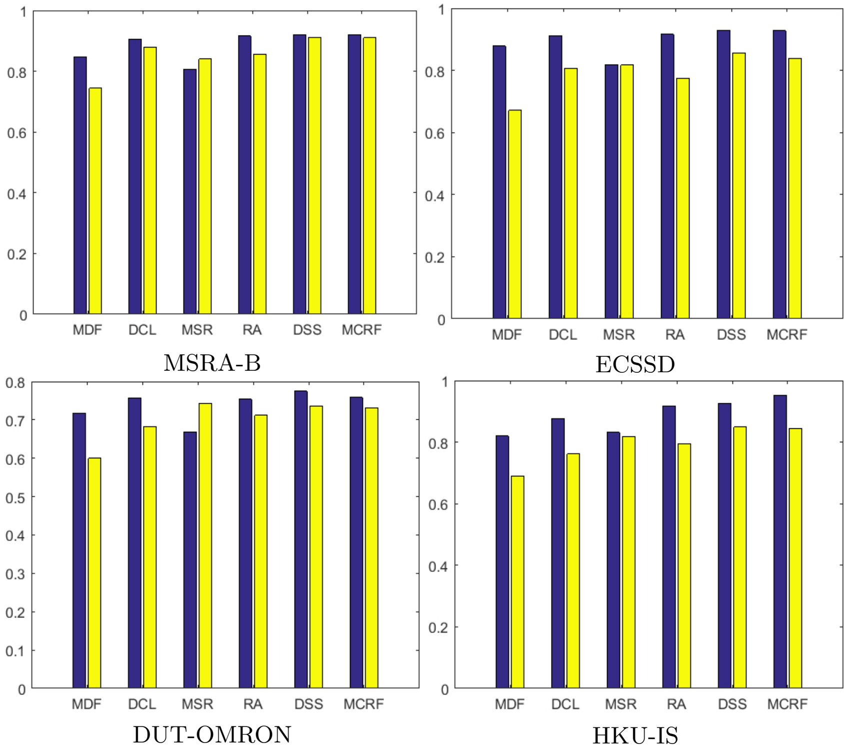

Fig. 2. Comparisons of the mean precision and mean recall on MSRA-B testing set, DUT-OMRON, HKU-IS and ECSSD datasets respectively.

Table 1 lists the mean F-measure and MAE of the nine saliency models and the proposed MCRF model over four datasets. It is clearly observed that the MCRF model surpasses most of the existing saliency models with much better performances. Compared to MDF [15], DCL [16], MSR [14] that apply single CRF [11] layers, the multi-scale CRF model results in superior performances. Moreover, the proposed MCRF model receives comparable performances with the DSS [9] model. Compared to the DSS [9] that uses the enhanced HED architecture with five scales of side outputs (totally 53 convolutional and deconvolutional layers), the proposed MCRF model is based on the simple encoder-decoder architecture and only three scales of side outputs (totally 31 convolutional and deconvolutional layers) are fused for multi-scale integration. Thus, the multi-scale CRF structure is proved to be efficient. We also evaluate the performances of embedding multi-scale CRFs versus single-scale CRFs to the pre-finetuned model as in Table 2. Clearly, multi-scale CRFs model receives the best performances. Figure 3 presents saliency maps from the compared models and the proposed MCRF model.

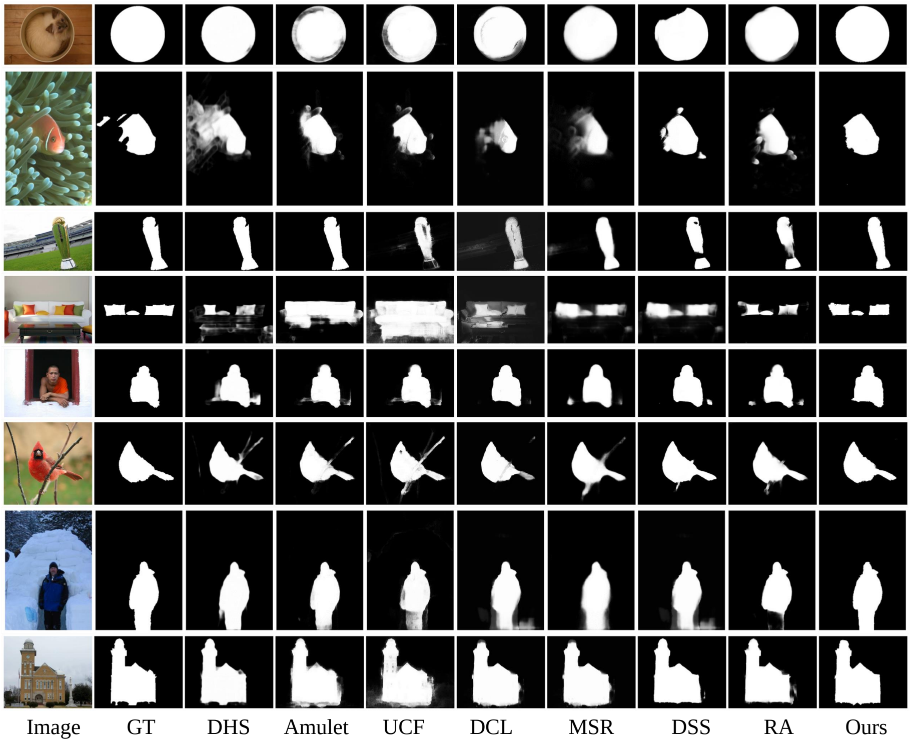

Fig. 3. Examples of saliency maps from DHS [18], Amulet [42], UCF [43], DCL [16], MSR [14], DSS [9], RA [4] and the proposed MCRF model.

5 Conclusion

This paper proposes to refine the side outputs efficiently from multiple scales of FCNs by embedding multi-scale CRF layers. Firstly, the front-end FCNs is based on the simple yet efficient encoder-decoder networks which involves much fewer convolutional layers and parameters such that the front-end network is easy to train. Secondly, only three scales of side outputs from the FCNs are integrated but competitive performances are received. In future, the side output refinement based on CRF inference with upper level side output from the FCNs will be further explored for a hierarchical refinement architecture.

References

1. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: Proc. CVPR. pp. 1597–1604. IEEE (2009)
2. Arbeláez, P., Pont-Tuset, J., Barron, J.T., Marques, F., Malik, J.: Multiscale combinatorial grouping. In: Proc. CVPR. pp. 328–335. IEEE (2014)

3. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. TPAMI **40**(4), 834–848 (2018)
4. Chen, S., Tan, X., Wang, B., Hu, X.: Reverse attention for salient object detection. In: Proc. ECCV. Springer (2018)
5. Dai, J., Li, Y., He, K., Sun, J.: R-fcn: Object detection via region-based fully convolutional networks. In: NIPS. pp. 379–387 (2016)
6. Guo, C., Zhang, L.: A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. TIP **19**(1), 185–198 (2010)
7. Han, J., Pauwels, E.J., De Zeeuw, P.: Fast saliency-aware multi-modality image fusion. Neurocomputing **111**, 70–80 (2013)
8. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Hypercolumns for object segmentation and fine-grained localization. In: Proc. CVPR. pp. 447–456. IEEE (2015)
9. Hou, Q., Cheng, M.M., Hu, X.W., Borji, A., Tu, Z., Torr, P.: Deeply supervised salient object detection with short connections. TPAMI (2018)
10. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proc. Multimedia. pp. 675–678. ACM (2014)
11. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: Advances in neural information processing systems. pp. 109–117 (2011)
12. Kuen, J., Wang, Z., Wang, G.: Recurrent attentional networks for saliency detection. In: Proc. CVPR. IEEE (2016)
13. Lee, G., Tai, Y.W., Kim, J.: Deep saliency with encoded low level distance map and high level features. In: Proc. CVPR. pp. 660–668. IEEE (2016)
14. Li, G., Xie, Y., Lin, L., Yu, Y.: Instance-level salient object segmentation. In: Proc. CVPR. pp. 247–256. IEEE (2017)
15. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: Proc. CVPR. IEEE (2015)
16. Li, G., Yu, Y.: Deep contrast learning for salient object detection. In: Proc. CVPR. IEEE (2016)
17. Li, G., Yu, Y.: Visual saliency detection based on multiscale deep cnn features. TIP **25**(11), 5012–5024 (2016)
18. Liu, N., Han, J.: Dhsnet: Deep hierarchical saliency network for salient object detection. In: Proc. CVPR. pp. 678–686 (2016)
19. Liu, Q., Hong, X., Zou, B., Chen, J., Chen, Z., Zhao, G.: Hierarchical contour closure based holistic salient object detection. TIP (2017)
20. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., Shum, H.Y.: Learning to detect a salient object. TPAMI **33**(2), 353–367 (2011)
21. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proc. CVPR. pp. 3431–3440. IEEE (2015)
22. Luo, P., Wang, X., Tang, X.: Pedestrian parsing via deep decompositional network. In: Proc. ICCV. pp. 2648–2655. IEEE (2013)
23. Ma, Y.F., Lu, L., Zhang, H.J., Li, M.: A user attention model for video summarization. In: Proc. Multimedia. pp. 533–542. ACM (2002)
24. Mahendran, A., Vedaldi, A.: Salient deconvolutional networks. In: Proc. ECCV. pp. 120–135. Springer (2016)
25. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: Proc. CVPR. pp. 891–898 (2014)

26. Qiu, W., Gao, X., Han, B.: A superpixel-based crf saliency detection approach. Neurocomputing **244**, 19–32 (2017)
27. Rahtu, E., Kannala, J., Salo, M., Heikkilä, J.: Segmenting salient objects from images and videos. In: Proc. ECCV. pp. 366–379. Springer (2010)
28. Santella, A., Agrawala, M., DeCarlo, D., Salesin, D., Cohen, M.: Gaze-based interaction for semi-automatic photo cropping. In: Proc. CHI. pp. 771–780. ACM (2006)
29. Sermanet, P., Kavukcuoglu, K., Chintala, S., LeCun, Y.: Pedestrian detection with unsupervised multi-stage feature learning. In: Proc. CVPR. pp. 3626–3633. IEEE (2013)
30. Shengfeng, H., Jianbo, J., ZHANG, X., Han, G., Rynson, W.: Delving into salient object subitizing and detection. In: Proc. ICCV. IEEE (2017)
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. ICLR (2015)
32. Tang, Y., Wu, X.: Saliency detection via combining region-level and pixel-level predictions with cnns. In: Proc. ECCV. pp. 809–825. Springer (2016)
33. Wang, L., Lu, H., Ruan, X., Yang, M.H.: Deep networks for saliency detection via local estimation and global search. In: Proc. CVPR. pp. 3183–3192. IEEE (2015)
34. Wang, L., Wang, L., Lu, H., Zhang, P., Ruan, X.: Saliency detection with recurrent fully convolutional networks. In: Proc. ECCV. pp. 825–841. Springer (2016)
35. Xie, S., Tu, Z.: Holistically-nested edge detection. In: Proc. ICCV. pp. 1395–1403. IEEE (2015)
36. Xu, D., Ouyang, W., Alameda-Pineda, X., Ricci, E., Wang, X., Sebe, N.: Learning deep structured multi-scale features using attention-gated crfs for contour prediction. In: NIPS. pp. 3964–3973 (2017)
37. Xu, D., Ricci, E., Ouyang, W., Wang, X., Sebe, N.: Monocular depth estimation using multi-scale continuous crfs as sequential deep networks. TPAMI (2018)
38. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: Proc. CVPR. pp. 1155–1162. IEEE (2013)
39. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: Proc. CVPR. pp. 3166–3173. IEEE (2013)
40. Yang, J., Price, B., Cohen, S., Lee, H., Yang, M.H.: Object contour detection with a fully convolutional encoder-decoder network. In: Proc. CVPR. pp. 193–202 (2016)
41. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Proc. ECCV. pp. 818–833. Springer (2014)
42. Zhang, P., Wang, D., Lu, H., Wang, H., Ruan, X.: Amulet: Aggregating multi-level convolutional features for salient object detection. In: Proc. ICCV. IEEE (2017)
43. Zhang, P., Wang, D., Lu, H., Wang, H., Yin, B.: Learning uncertain convolutional features for accurate saliency detection. In: Proc. ICCV. pp. 212–221. IEEE (2017)
44. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.: Conditional random fields as recurrent neural networks. In: Proc. ICCV. pp. 1529–1537. IEEE (2015)