Semantical Knowledge Guided Salient Object Detection with Multiple Proposals

Xue Zhang[1], Zheng Wang*[1], and Meijun Sun[1]

Tianjin University, Tianjin 300350, China
Dorothyzhx@foxmail.com, wzheng@tju.edu.cn, sunmeijun@tju.edu.cn

Abstract. In recent years, the area of salient object detection has developed rapidly duo to the revival of deep learning techniques, especially the emergence of deep Convolutional Neural Networks(CNNs), which has greatly boosted the detection result. Although CNNs can be used to perceive the salient objects, it is difficult to work for images with complex background, which is also a major problem faced by most work. In this paper, we introduce a salient object detection method using high-level features with semantic meaning. In order to obtain the accurate region and edge of all the salient objects in an image, our model has two designs: (1) utilizing multiple proposals as the semantical knowledge prior to enhance the power of locate objects, that are most likely to cover the entire salient regions of an image, and (2) using several attention modules to improve the representation ability of our model ,and using abundant low-level feature information extracted by the encoder of the network to assist its decoder to obtain the precise saliency map relatively. In addition, our model can make full use of multi-level features and semantical knowledge, so the saliency map we got is very excellent. The experiments shows that our approach achieves state-of-the-art performance on four public benchmarks, and produces significant improvements over existing well-known methods.

Keywords: Salient object detection · Object proposals · Semantical knowledge.

1 Introduction

Salient object detection(SOD) aims to highlight the most conspicuous regions in images. With the development of deep learning, SOD can be greatly applied to other tasks as a preprocessing procedure, such as image captioning [3,16], image matching [30,31], visual question answering [14], object detection [6,12], person re-identification [36], *etc.*

In the past decades, SOD technology has been dramatically evolved. Previous SOD methods utilize hand-craft visual features [27,37] and heuristic priors [1,32] to detect salient objects. However, most of them focus on low-level features and can not fully catch high-level semantic features, so that the saliency maps got

* Z. Wang is the corresponding author.

by them are not to our satisfactory. Recently, the emergence of deep CNNs has opened the door of the rapid development of SOD. Deep CNNs [22] extract image features in an end-to-end manner and then get salient objects after training the model. Convolution operations near the bottom of the network can obtain low-level feature information. With the deepening of the network, high-level semantic information becomes stronger. Therefore, deep CNNs can automatically locate the salient objects more accurate than most detection methods used before the age of deep CNNs.

In order to make full use of the low-level information which is sensitive to edges and the high-level semantic information which is sensitive to salient objects, we extract the feature maps from the encoder of the network and fuse them into the decoding stage, so that the model can make the best of the different level features in every stage. What's more, we also import several attention modules to make full use of the spatial and channel features transferring between the adjacent layers. More importantly, for the purpose of perceiving the salient objects that are most likely cover all the salient regions quickly and accurately, we use multiple object proposals as semantical knowledge prior to guide the model. Our experimental results clearly demonstrate the effectiveness of the proposed method on four well-known benchmark datasets.

In summary, the main contributions of our work are as follows:

- We propose a method to detect salient objects in images using multiple proposals as semantical knowledge. With rich information of semantics about proposals, the model can perceive all salient regions faster and better.
- Our model makes the most of the multi-level features extracted at every stage of the model by integrating the low-level features got at the bottom of the model with the high-level features got at the top of the model. In addition, we are the first to use attention modules in the saliency detection, and we put several attention modules into the model to optimize features transferring between layers to get a better result.
- Our method shows good generalization and yields comparable even better performance than the state-of-the-art SOD methods on several public datasets.

2 Related Work

SOD technology has attracted lots of interests among the computer vision scholars. Unlike fixation [4], one of the attention mechanism work, SOD aims at extracting all the salient regions with clear contour and represents them using a binary graph. SOD methods can be mainly divided into hand-craft features based approach and learning based approach according to the way to get features of images. For the hand-crafted methods [9, 29], they can be traced back to [20], Treisman proposed a feature-integration model of attention. The model selects the most important visual features and combines them to get the salient objects. Wang *et al.* [21] treated the saliency computation as the regression problem. Lee *et al.* [11] proved that the hand-crafted features can provide additional

information to boost the performance of saliency detection, which depends on high-level features only. Xie *et al.* [26] proposed a new computational saliency detection model which is implemented with a coarse to fine strategy under the Bayesian framework. Although the hand-craft SOD methods can achieve a nice detection, they are not robust enough to deal with the complex scenarios. Thus, more and more researcher pay attention to learning based approach [13, 22] to bring an accurate result. Hou *et al.* [7] introduced a series of short connections between deeper and shallower layers, and the activation of each layer can highlight the corresponding objects and detect their boundaries accurately. Liu *et al.* [15] solved the SOD problem by expanding the pooling part of the convolutional neural network. Zhang *et al.* [35] integrated multi-level feature maps into multiple resolution at first, which includes rough semantics and fine details and then learned to combine these feature maps at each resolution adaptively. Finally, the results were fused effectively to generate the final saliency map. Particularly, Guo *et al.* [5] utilized the object porposals to detect salient objects, and they considered the SOD as the ranking and voting strategy to object proposals. Different from using the object proposals to make up the saliency map, we use multiple object proposals as semantical knowledge to assist the model in perceiving the most conspicuous regions in images quickly and exactly.

3 Approach

3.1 Overview

In this work, we describe a procedure for building and learning deep image saliency detection networks using the multiple objects proposals as high-level prior. An overview of our method is presented in Fig. 1. From the macro point of view, the process for our method to detect salient objects can be divided into to two main steps roughly. Firstly, we put an image I into the proposal extractor and get a lot of object proposals. Then we calculate a score for each object proposal and order them according to the score. We chose top 10 as the final proposals used in our model which are most likely to cover all the salient regions of an image. Secondly, the 10 object proposals P are thrown into the network to catch the saliency map F after the encoding-decoding operations. The training of our model can be achieved by minimizing the following objective function:

$$\min_{\theta} \sum_{k} Loss(L(I_k), F_k(I_k; \theta)) \tag{1}$$

where $L(I_k)$ represents the label of the *k-th* image I_k, and F_k is the saliency map got by our method. θ denotes the model parameter. $Loss(\cdot)$ is the per-pixel loss function.

In the rest of this section, we explain the procedure of data preprocessing and the prediction network in detail.

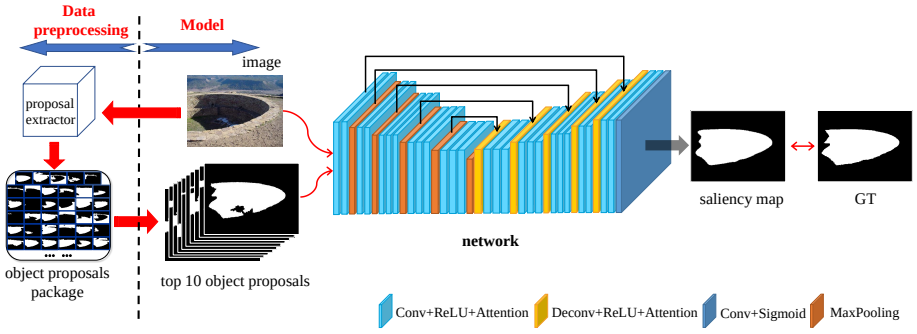

Fig. 1. The overview of our approach. To understand the work procedure better, we also place the data preprocessing pipeline on the left of the figure.

3.2 Data preprocessing

The multiple proposals used in our method are obtained from the proposal extractor [8]. Given an image I_i, the extractor can produce hundreds of object proposals. Here we represent them with P. However, not all of these proposals are useful for our methods, because lots of them are mixed with much noise strongly. The working principle of the proposal extractor is to generate a set of segmentations by performing graph cuts, and rank them according to their importance. To get the object proposals which are most likely to cover all the salient regions, we rank them again by computing the new Intersection-over-Union(IoU) score τ, which are modified by our method, between all of the extracted proposals and the label of image. Then we select top 10 of them with the score greater than 3 as the high-level semantic knowledge experimentally. The new IoU score can be written as Eq (2):

$$\tau = \frac{P_j(I_i) \bigcap L(I_i)}{P_j(I_i) \bigcup L(I_i) - P_j(I_i) \bigcap L(I_i)} \tag{2}$$

where $P_j(I_i)$ is the j-*th* proposal got by the extractor for the i-*th* image I_i.

3.3 Network Structure

Our network is rooted in VGG-16 model [18], which is pre-trained on the ImageNet, but we have improved it by adding several attention modules proposed in [25] between adjacent convolutional layers. We also integrate the low-level features of the encoder into the high-level features of the decoder by concatenating each pair of them. What needs to be added is that the attention module has two sequential sub-modules, channel attention module and spatial attention module. Because of the outstanding performance of attention modules for feed-forward convolutional networks, we use them behind every convolutional layer to transfer features more efficiently. Thus, the useful features can be fully utilized to

perceive the salient regions. Moreover, we know that the convolution layers near the bottom are more sensitive to low-level features, while the deeper convolutional layers are easier to perceive semantic information. We integrate the *m-th* low-level features $f_{I_i}^{m_{low}}$ and the *m-th* high-level features $f_{I_i}^{m_{high}}$ in the network in order to get a more accurate saliency map. Therefore, the fused feature map $\tilde{f}_{I_i}^{m}$ for the image I_i is formally written as:

$$\tilde{f}_{I_i}^{m} = \sigma(Concat(f_{I_i}^{m_{low}}, f_{I_i}^{m_{high}}) \otimes w_{I_i}^{m} + b_{I_i}^{m}) \tag{3}$$

where $w_{I_i}^{m}$ represents the filters for the *m-th* de-convolution operation for the image I_i, and b_i^m is its biases. $Concat(\cdot)$ is the concatenate operation between the two features, and $\otimes$ represents convolution operation. $\sigma(\cdot)$ refers to ReLU, a kind of non-linear activation function.

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation Metrcis: We evaluate our method on four public benchmark datasets. **HKU-IS** [17] consists of 4,447 images including multiple disconnected objects. **ECSSD** [28] has 1,000 natural images with complex structures. **PASCAL-S** [33] comes from the PASCAL VOC [2] and contains 850 images. **MSRA5K** [19] has 5,000 images of all kinds.

We evaluate the performance of our model and compare with other works by four widely used metrics, i.e., Precision-Recall Curve(PR Curve), Area Under Curve(AUC) score, F-measure and Mean Absolute Error(MAE). F-measure, which we denotes as F_β, is an overall performance measure. It can be computed by the weighted harmonic mean of the precision and recall:

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{4}$$

where β^2 is set to 0.3 following [15]. MAE is the similarity between a saliency map S and ground truth G:

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |S(x,y) - G(x,y)| \tag{5}$$

where S is the saliency map and G represents the ground truth. W and H denote the width and height of S respectively.

Implementation: Our experiments are conducted with a NVIDIA GTX 1080Ti GPU. The code is based on Python with the Keras toolbox. Our method is rooted in the VGG-16 model and parameters in other layers are initialized randomly. In the training stage, all the images are resized to 224×224 and we use Adam optimizer and the learning rate is set to 10^{-6}.

4.2 Performance Comparison

We compare our method with three state-of-the-art deep learning based SOD approaches including RFCN [23], C2SNet [13], WSS [22], and five conventional counterparts including HS [28], MBS [34], SF [10], wCtr [37] and GS [24]. As shown in Fig.2 and Table.1, our proposed method outperform existing works across almost all the datasets according to the evaluation metrics, and they demonstrate the effectiveness of our proposed method strongly.

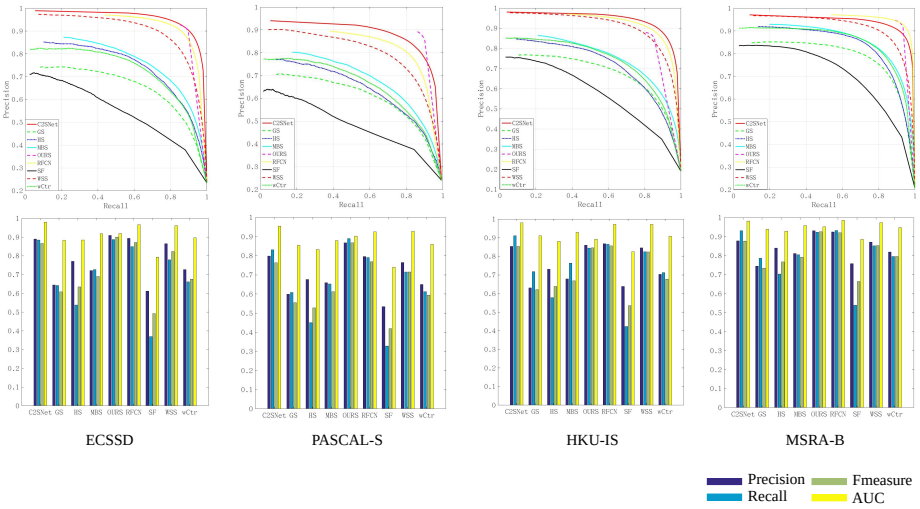

Fig. 2. The PR curves and bar graphs show the comparison of nine saliency maps on four popular salient object datasets.

Table 1. Quantitative evaluations. The best three scores are shown in red, blue, green, respectively

Methods	MSRA-B		PASCAL-S		HKU-IS		ECSSD	
	F-measure	MAE	F-measure	MAE	F-measure	MAE	F-measure	MAE
HS	0.7669	0.1621	0.5272	0.2625	0.6377	0.2150	0.6347	0.2274
MBS	0.7922	0.1116	0.6126	0.1964	0.6678	0.1503	0.6903	0.1707
SF	0.6640	0.1660	0.4187	0.2358	0.5347	0.1744	0.4921	0.2187
wCtr	0.7959	0.1106	0.5935	0.1986	0.6770	0.1424	0.6762	0.1712
GS	0.7348	0.1445	0.5553	0.2209	0.6213	0.1681	0.6080	0.2058
RFCN	0.9211	0.0346	0.7685	0.1036	0.8564	0.0546	0.8714	0.0667
C2SNet	0.8765	0.0478	0.7632	0.0805	0.8534	0.0460	0.8666	0.0535
WSS	0.8535	0.0763	0.7151	0.1395	0.8237	0.0790	0.8233	0.1039
OURS	0.9267	0.0287	0.8678	0.0544	0.8455	0.0555	0.9002	0.0451

From the performance comparison with the state-of-the-art, we can see that our method can largely outperform other leading methods on MSRA-B, PASCAL-S and ECSSD. Particularly, our method increases the highest F-measure score by 0.61%, 12.92%, 3.31% and decreases the lowest MAE score by 17.05%, 32.42%, 15.70%, respectively.

We also visualize some example saliency maps of our model in Fig.3. From the picture we can see that our method can achieve more accurate results.

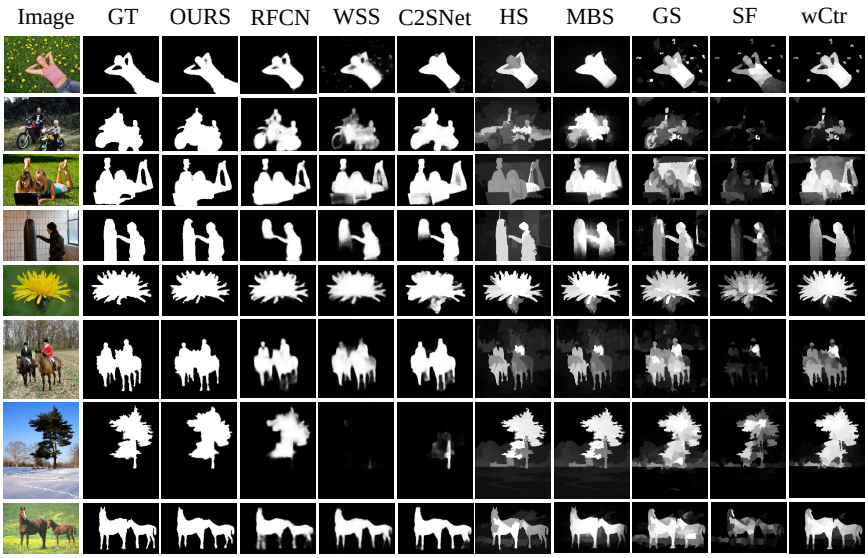

Fig. 3. Qualitative comparisons to previous state-of-the-art methods. The rows 1 shows the object touching the boundary of the image. The rows 2 and 6 show the low contrast between objects and backgrounds. The rows 2 - 4 and 6, 8 show the comparison in multiple objects. The rows 5 and 7 show the objects with complicated edges.

5 Conclusions

In this paper, we propose a new method to detect salient objects by using multiple proposals as semantical knowledge. We first use the proposal extractor to acquire hundreds of object proposals, and then rank them using the method we mentioned earlier. To avoid too much computation, we chose top 10 of them as the most semantical knowledge. Moreover, we are the first to insert attention modules into VGG-16 model to detect salient objects, and we also mix the low-level features of the encoder with the high-level features of the decoder to get a much better saliency map. Extensive evaluations demonstrate that our method

can improve the performance of SOD significantly and show nice generalization of our model.

References

1. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H., Hu, S.M.: Global contrast based salient region detection. In: Computer Vision and Pattern Recognition (2011)
2. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision **88**(2), 303–338 (Jun 2010)
3. Fang, H., Gupta, S., Iandola, F., Srivastava, R., Zweig, G.: From captions to visual concepts and back. In: IEEE Conference on Computer Vision and Pattern Recognition (2014)
4. Gorji, S., Clark, J.J.: Attentional push: A deep convolutional network for augmenting image salience with shared attention modeling in social scenes. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
5. Guo, F., Wang, W., Shen, J., Shao, L., Yang, J., Tao, D., Tang, Y.Y.: Video saliency detection using object proposals. IEEE Transactions on Cybernetics **PP**(99), 1–12 (2017)
6. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Torr, P.: Deeply supervised salient object detection with short connections. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
7. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.H.S.: Deeply supervised salient object detection with short connections. IEEE Conference on Computer Vision and Pattern Recognition pp. 5300–5309 (2017)
8. Ian Endres, D.H.: D.: Category independent object proposals. In: Europeon Conference on Computer Vision (2010)
9. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. Pattern Analysis and Machine Intelligence, IEEE Transactions on **20**, 1254 – 1259 (12 1998). https://doi.org/10.1109/34.730558
10. Krahenbuhl, P.: Saliency filters: Contrast based filtering for salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition (2012)
11. 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 (2016)
12. Li, G., Yu, Y.: Deep contrast learning for salient object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
13. Li, X., Yang, F., Cheng, H., Liu, W., Shen, D.: Contour knowledge transfer for salient object detection. In: Europeon Conference on Computer Vision (2018)
14. Lin, Y., Pang, Z., Wang, D., Zhuang, Y.: Task-driven visual saliency and attention-based visual question answering. CoRR **abs/1702.06700** (2017)
15. Liu, J., Hou, Q., Cheng, M.M., Feng, J., Jiang, J.: A simple pooling-based design for real-time salient object detection. CoRR **abs/1904.09569** (2019)
16. Ramanishka, V., Das, A., Zhang, J., Saenko, K.: Top-down visual saliency guided by captions. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
17. Rui, Z., Ouyang, W., Li, H., Wang, X.: Saliency detection by multi-context deep learning. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)

18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2015)
19. Tie, L., Zejian, Y., Jian, S., Jingdong, W., Nanning, Z., Xiaoou, T., Heung-Yeung, S.: Learning to detect a salient object. IEEE Trans Pattern Anal Mach Intell **33**(2), 353–367 (2011)
20. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. Cognitive Psychology **12**(1), 97–136 (1980)
21. Wang, J., Jiang, H., Yuan, Z., Cheng, M.M., Hu, X., Zheng, N.: Salient object detection: A discriminative regional feature integration approach. International Journal of Computer Vision **123**(2), 251–268 (2017)
22. Wang, L., Lu, H., Wang, Y., Feng, M., Xiang, R.: Learning to detect salient objects with image-level supervision. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
23. Wang, L., Wang, L., Lu, H., Zhang, P., Xiang, R.: Saliency detection with recurrent fully convolutional networks. In: European Conference on Computer Vision (2016)
24. Wei, Y., Fang, W., Zhu, W., Jian, S.: Geodesic saliency using background priors. In: European Conference on Computer Vision (2012)
25. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Europeon Conference on Computer Vision (2018)
26. Xie, Y., Lu, H.: Visual saliency detection based on bayesian model. In: IEEE International Conference on Image Processing (2011)
27. Xin, L., Fan, Y., Chen, L., Cai, H.: Saliency transfer: an example-based method for salient object detection. In: International Joint Conference on Artificial Intelligence (2016)
28. Yan, Q., Li, X., Shi, J., Jia, J.: Hierarchical saliency detection (2013)
29. Yang, C., Zhang, L., Lu, H., Xiang, R., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: IEEE Conference on Computer Vision and Pattern Recognition (2013)
30. Yang, T.Y., Hsu, J.H., Lin, Y.Y., Chuang, Y.Y.: Deepcd: Learning deep complementary descriptors for patch representations. In: IEEE International Conference on Computer Vision (2017)
31. Yang, T.Y., Lin, Y.Y., Chuang, Y.Y.: Accumulated stability voting: A robust descriptor from descriptors of multiple scales. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
32. Yao, Q., Lu, H., Xu, Y., He, W.: Saliency detection via cellular automata. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)
33. Yin, L., Hou, X., Koch, C., Rehg, J.M., Yuille, A.L.: The secrets of salient object segmentation (2014)
34. Zhang, J., Sclaroff, S., Zhe, L., Shen, X., Price, B., Mech, R.: Minimum barrier salient object detection at 80 fps. In: IEEE International Conference on Computer Vision (2015)
35. Zhang, P., Wang, D.K., Lu, H., Wang, H., Ruan, X.: Amulet: Aggregating multi-level convolutional features for salient object detection. IEEE International Conference on Computer Vision pp. 202–211 (2017)
36. Zhao, R., Ouyang, W., Wang, X.: Person re-identification by saliency learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(2), 356–370 (2017)
37. Zhu, W., Shuang, L., Wei, Y., Jian, S.: Saliency optimization from robust background detection. In: IEEE Conference on Computer Vision and Pattern Recognition (2014)