MULTI-PATH FEATURE FUSION NETWORK FOR SALIENCY DETECTION

Hengliang Zhu[1]**, Xin Tan*[1]*, Zhiwen Shao*[1]*, Yangyang Hao*[1]*, Lizhuang Ma*[1,2]***

[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
[2]Department of Computer Science and Software Engineering, East China Normal University, China
{hengliang_zhu, tanxin2017, shaozhiwen, haoyangyang2014}@sjtu.edu.cn, ma-lz@cs.sjtu.edu.cn

ABSTRACT

Recent saliency detection methods have made great progress with the fully convolutional network. However, we find that the saliency maps are usually coarse and fuzzy, especially near the boundary of salient object. To deal with this problem, in this paper, we exploit a multi-path feature fusion model for saliency detection. The proposed model is a fully convolutional network with raw images as input and saliency maps as output. In particular, we propose a multi-path fusion strategy for deriving the intrinsic features of salient objects. The structure has the ability of capturing the low-level visual features and generating the boundary-preserving saliency maps. Moreover, a coupled structure module is proposed in our model, which helps to explore the high-level semantic properties of salient objects. Extensive experiments on four public benchmarks indicate that our saliency model is effective and outperforms state-of-the-art methods.

Index Terms— Multi-path feature fusion, coupled structure, saliency detection, fully convolutional network

1. INTRODUCTION

Saliency detection aims at locating the most visually distinctive regions in an image, and has attracted many researchers in recent years. From the perspective of biological visual perception, salient object has abundant visual information, which is first attracted by human attentions. In many computer vision tasks, saliency detection is usually utilized as a preprecessing step. It has been shown a great success in many visual fields, such as image segmentation [1], object tracking [2] and content-aware image resizing [3].

Inspired by the human visual attention mechanisms, numerous saliency algorithms have been proposed to detect the conspicuous object from an image. Traditional algorithms [4–7] are essentially based upon low-level features, such as color, texture and location cues. These algorithms usually exploit lots of prior knowledge to detect salient objects, including local or global contrast, center-surround difference and boundary connectivity [8]. Though obvious progress has been made, the main drawbacks of these algorithms are that the saliency results are not smooth at the object boundary, or fail to extract the complete salient object.

Recently, many deep learning based algorithms [9–13] have been proposed for saliency detection. These algorithms use the convolutional neural networks (CNNs) to predict the salient object in an end-to-end manner, and reach a remarkable performance in terms of accuracy. The CNNs algorithms are effective for detectting salient object, but the saliency maps are still unsatisfactory. Due to lacking of the low-level visual information, the saliency maps are usually blurry, especially at the objects boundaries. On one hand, these algorithms simply integrate the high-level semantic features to generate the final results. In fact, the natural images usually contain diverse structures (i.e. clutter texture), so it is challenging to preserve the contour of the salient object. On the other hand, when background regions have similar appearances as the foreground, the proposed algorithms are not able to highlight the whole object uniformly.

Based on above discussions, we pay attention to the two crucial problems. First, how to efficiently utilize the multi-level convolutional features and preserve the objects' structures information (such as edges). Second, how to precisely locate the salient regions and uniformly highlight the saliency maps. To deal with this two challenges simultaneously, in this paper, we propose a novel multi-path feature fusion networks (see Fig.1) to detect the salient objects. Similar to the state-of-the-art algorithms [12, 13], the proposed saliency model is also based on a fully convolutional neural network with the raw images as input and the whole saliency maps as output. In our network, a simple yet efficient fusion method is used to utilize the multi-level features. Moreover, inspired by DenseNet [14], we propose a coupled structure module (namely CSM) to reuse the convolution features. The structure can improves the accuracy of saliency detection and generate the highlighted saliency map. Experimental results show that our saliency model has the capability of capturing the rich features of salient object across different convolutional layers.

Our study attempts to design an efficient deep convolu-

*Corresponding author. This work is supported by the National Natural Science Foundation of China (No. 61472245 and 61502220), and the Science and Technology Commission of Shanghai Municipality Program (No. 16511101300).

978-1-5386-1737-3/18/$31.00 ©2018 IEEE

tional network for saliency detection in complex images. The main contributions are summarized as follows:

- We propose a multi-path feature fusion network, namely MPFFNet, which uses the convolutional features from different paths to generate the high-quality saliency maps. The MPFFNet can extracts the intrinsic properties of salient object with combining the multi-level features.
- We also propose a coupled structure module to improve the accuracy of predicting the target object. Benefiting from this module, the performance of saliency detection is greatly improved for better salient object localization.
- Our algorithm can uniformly highlights the salient objects and smooths the saliency values in the salient objects boundaries. Comprehensive experiments show that our algorithm achieves competitive performance compared with state-of-the-art algorithms.

2. RELATED WORK

In this section, we mainly review the classic deep learning methods for saliency detection. In recent years, numerous superior models have been proposed for saliency detection by using the convolutional neural networks. These methods make attempts to develop various deep architectures for capturing the intrinsic features of salient objects.

For instance, Zhao et al. [11] designed a multi-context deep learning framework for saliency detection. They utilized global and local context information to model the visual saliency. Li et al. [9] used multi-scale features extracted from a deep CNNs to produce a high-quality saliency map. Lee et al. [10] proposed a unified deep learning framework for salient object detection by integrating both high level and low level semantic properties. Wang et al. [15] detected salient object by integrating both local estimation and global search. Two different deep architectures are used to derive the rich features of salient object. However, these methods could not predict the object details ideally.

Recently, saliency detection and semantic segmentation are mostly based on fully convolutional networks (FCNs), and have reached the outstanding results. Liu et al. [12] proposed an end-to-end deep hierarchical network for saliency detection. They first predicted salient maps coarsely by using a global information, then progressively improved object details by integrating the local context. Li et al. [13] proposed a deep contrast network for detecting salient object, which includes a pixel-level fully convolutional stream and a segment-wise spatial pooling stream. Wang et al. [16] utilized a recurrent fully convolutional networks for salient object detection. Zhang et al. [17] learned the deep uncertain convolutional features for saliency detection, which introduced a reformulated dropout after specific convolutional layers. Li et al. [18] proposed a multi-task deep model based on a fully convolutional network. Though great progress have been made by these saliency models, there is still a large room for exploiting multi-level convolutional features that preserve salient object boundary and structure.

3. THE PROPOSED METHOD

In this section, we illustrate the proposed algorithm in details. As shown in Fig. 1, the architecture of our deep saliency model consists of two components, a multi-path feature fusion framework and a coupled structure module. The multipath fusion is a spatial pyramid structure, which is used to produce different semantic feature maps. For each path, we use a simple way to integrate its feature map with shallow convolutional layer. At the same time, we propose a coupled structure module to improve the accuracy of localization for saliency detection. Finally, the final saliency map is generated by aggregating maps from different paths.

3.1. Multi-path feature fusion

In order to efficiently use multi-level convolutional features, we design a multi-path structure to fuse the features at different convolutional layers. Our saliency algorithm is based on the VGG16 model [19], which has a strong feature representation ability. The proposed network has five max-pooling layers with kernel size 2 and stride 2, so the size of output feature map is reduced by a factor of 32. To make the output map have the same size as input image, we use the upsampling operation to scale the map. In order to improve the accuracy of saliency detection, we add three paths after the last pooling stage to build the spatial pyramid structure. As shown in Fig. 1, the output of the path module (yellow boxes, i.e. Path1, Path2, Path3) represents high-level feature map for each branch, denoted as P_i. Fig. 2 (right) shows that each path module includes one 3×3 convolutional layer, one coupled structure module, one 1×1 convolutional layer and one up-sampling layer. Moreover, we observe that the multi-path structure can accurately localize the salient regions but lose lots of details. Previous work [20] indicates that the low-level

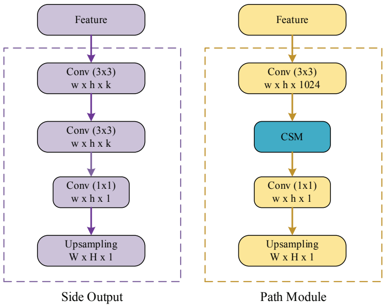


Fig. 2. The structure of side output and path module.

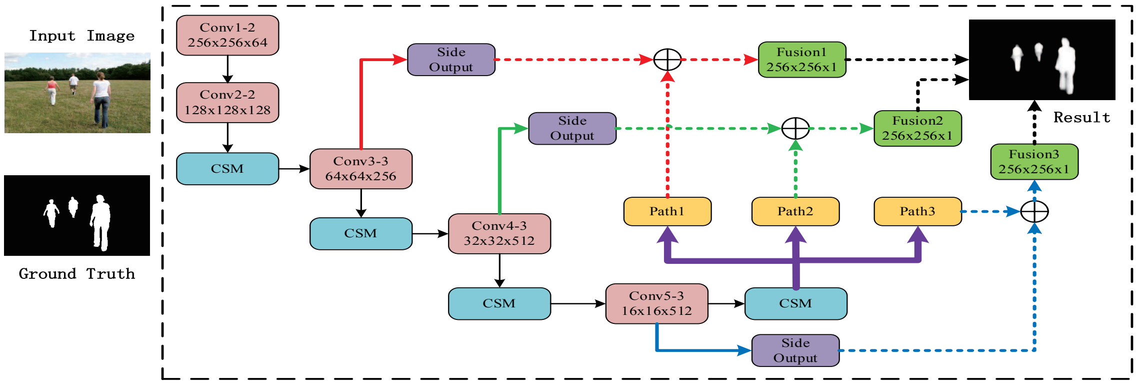


Fig. 1. The architecture of our saliency detection model (best viewed in color). Similar to previous CNN-based methods [13, 17], the proposed end-to-end network is also built on VGG16 model. The model has three primary parts: (1) blue box (CSM) is a coupled structure module. (2) yellow boxes (i.e. $Path_i$) represent feature extraction function for each branch. (3) three extra branches are connected with shallow convolutional layers, denoted as purple box (Side Output). Given an original image (256x256x3), the high-level feature of salient object is first extracted by each path. Then three fusion maps are obtained by integrating these features. At last, the final saliency result is generated by merging the three feature maps.

visual information is conducive to improve object details, especially in salient object boundary. Therefore, we connect three extra branches with shallow convolutional layers. These shallow layers are conv3-3, conv4-3 and conv5-3, respectively. Each branch contains two 3×3 convolutional layers, one 1×1 convolutional layer and one up-sampling layer, as shown in Fig. 2 (left). Then, three output feature maps (Side Output) that represent rich contextual information are generated by each branch, denoted as B_i. These hierarchical feature maps $\{B_i, i = 1, 2, 3\}$ contain abundantly low-level visual properties that is complementary to the high-level feature maps $\{P_i, i = 1, 2, 3\}$. Thus, we further fuse these feature maps, defined as

$$F_i = W_i * Concat(P_i, B_i), \tag{1}$$

where the symbol $*$ is the convolution operation; $Concat$ is the cross-channel concatenation. W_i is a convolutional layer with 1×1 kernel size, which is used to balance the importance of each feature map. For the final prediction, the feature map is written as

$$S = V * Concat(F_1, F_2, F_3), \tag{2}$$

where V is also a 1×1 convolutional layer. We use the cross entropy in our model, and compute the loss function between the fused saliency map and the ground truth. Given a input image X, we define its corresponding saliency map as $Y = \{y_i, i = 1, ..., |Y|\}, y_i \in [0, 1]$. Thus, the final loss function can be represented by

$$\begin{aligned} L_{final} = - \sum_{y_i \in Z} & y_i log P(y_i = 1 | X; \Phi) \\ & + (1 - y_i) log P(y_i = 0 | X; \Phi). \end{aligned} \tag{3}$$

The parameter Φ is the collection of all network weights, which are updated using SGD algorithm at each iteration. Experimental results show that the final feature map can significantly keep the details of salient object contour.

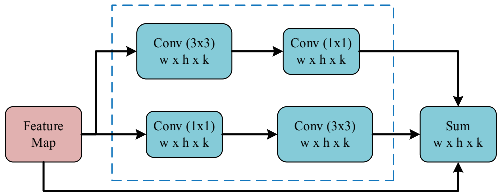


Fig. 3. The architecture of CSM.

3.2. Coupled structure module

In complex images, salient objects often have different patterns, such as size, shape and position, so the saliency model should be able to efficiently obtain the high-level features of salient object. Previous work [20] indicates that the high-level semantic feature can improve the accurate recognition of salient object. Therefore, the task of saliency detection needs to address the problems: how to capture the spatial information and accurately localize the salient object.

More recently, feature reuse is widely used in the modern deep networks, such as ResNet and DenseNet. For example, semantic segmentation has achieved outstanding results by using these technologies. Motivated by this achievements, we propose a coupled structure module embedded in the network to improve the performance of localization, as shown in Fig.3. The coupled structure consists of two complementary and symmetric components. Each component includes one 3×3 convolutional layer and one 1×1 convolutional layer. The non-linear transformation (ReLU) is also used after each convolutional layer. Besides, in order to enlarge the

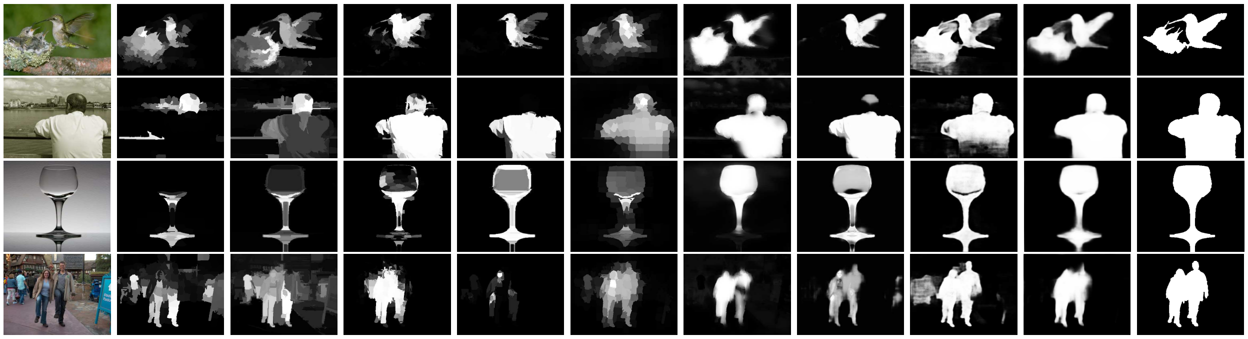


Fig. 4. Visual comparison of saliency maps generated from different saliency models.

receptive filed to cover the whole object, we use the dilated convolutions [21] to increase the size of filters. The coupled structure module is spatially sensitive and has the capability of extracting the high-level semantic features. Experimental results show that the CSM can conspicuously promote the accuracy of saliency detection.

4. EXPERIMENTS

In this section, we demonstrate the effectiveness of multi-path feature fusion model and show the performance of our algorithm compared with state-of-the-art algorithms. We also analyze the importance of the coupled structure module.

4.1. Datasets

For the training, we use MSRA10K dataset [22] to train the network without data augmentation, which includes $10,000$ images with pixel-wise annotations. This dataset is simple and contains only one salient object in an image. For the performance evaluation, we compare our model with state-of-the-art algorithms on four public benchmarks, including ECSSD [23], PASCAL-S [24], HKU-IS [11] and DUT-OMRON [25]. ECSSD is a very challenging dataset, it contains $1,000$ structurally complex images. PASCAL-S contains 850 images, which is selected from the PASCAL VOC 2012 dataset. HKU-IS is a large dataset that contains $4,447$ challenging images with low contrast or multiple salient objects. In our experiments, we use the testing images of HKU-IS for evaluation. DUT-OMRON has $5,168$ challenging images with complex backgrounds. Many images in this dataset contain one or more salient objects.

4.2. Evaluation metrics

We use three metrics to evaluate the performance of different saliency models, including the precision-recall (PR) curves, maximum F-measure and mean absolute error (MAE) [8]. The saliency map is first converted to a binary mask using a threshold within the range of [0, 255]. Then the PR curves are obtained by comparing the binary mask with the ground truth. The F-measure is defined as

$$F_\omega = \frac{(1+\omega^2) * precision * recall}{\omega^2 * precision + recall}, \tag{4}$$

where ω^2 is set to 0.3 like the most previous works [4, 6, 13]. We also report the MAE results, which measure the pixel-level difference between the saliency map and the ground truth.

4.3. Implementation Details

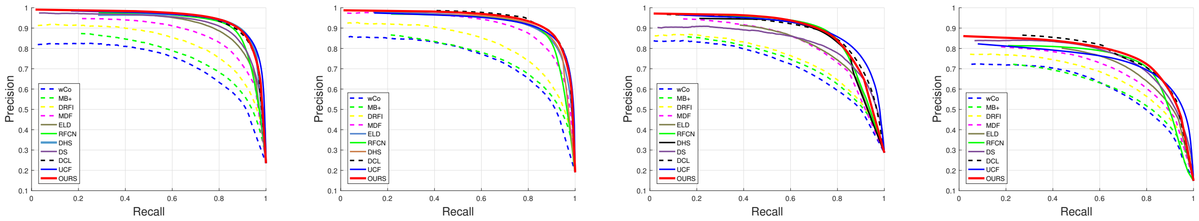
We train our MPFFNet with an open source deep learning framework Caffe [26], and directly feed the input images into the network. The proposed algorithm is trained on an Intel Core computer with an i7-6850K CPU and a single GeForce GTX 1080Ti GPU. In our experiments, we utilize VGG16 as our pre-trained model and set the base learning rate to $1e-8$. The parameter of momentum is set to 0.9 and the weight decay is set to 0.0005.

Running time. For training stage, it takes us about 10 hours to train the deep model. In testing, our network takes 0.056s (18 FPS) to process an image (average 400×300) without any pre/post-processing. Our method is faster than most existed convolutional methods, for example, MDF(8s) [9], DCL(1.5s) [13], UCF(0.14s) [17], ELD(0.5s) [10], DHS(0.04s) [12], and RFCN(4.6s) [16].

4.4. Comparison with other methods

We compare our algorithm with seven deep learning based algorithms, including UCF [17], DCL [13], DHS [12], DS [18], ELD [10], MDF [9], RFCN [16]. We also compare our algorithm with three classic algorithms: DRFI [6], wCo [4], and MB+ [5]. For a fair comparison, we use the saliency results or source code provided by the author.

PR curves. As shown in Fig. 5, we report the performance of PR curves compared with above-mentioned algorithms. Benefiting from the multi-path feature fusion, the



(a)ECSSD (b)HKU-IS (c)PASCAL (d)DUT_OMRON

Fig. 5. The performance of PR curves on four datasets.

Table 1. Quantitative comparison of different methods on four datasets. The best two results are shown in red and blue.

Method	ECSSD		HKU-IS		PASCAL-S		DUT-OMRON	
	maxF	MAE	maxF	MAE	maxF	MAE	maxF	MAE
wCo [4]	0.7156	0.1713	0.7255	0.1405	0.6550	0.1924	0.6298	0.1411
MB+ [5]	0.7389	0.1707	0.7328	0.1492	0.6765	0.1908	0.6242	0.1679
DRFI [6]	0.7860	0.1644	0.7826	0.1431	0.6898	0.1969	0.6640	0.1496
MDF [9]	0.8316	0.1050	0.8605	0.1292	0.7636	0.1453	0.6944	0.0916
ELD [10]	0.8681	0.0790	0.8809	0.0628	0.7713	0.1233	0.7052	0.0910
RFCN [16]	0.8976	0.0952	0.8876	0.0795	0.8320	0.1163	0.7381	0.0945
DS [18]	0.8824	0.1217	-	-	0.7601	0.1625	0.7449	0.1204
DCL [13]	0.9006	0.0747	0.9066	0.0552	0.8105	0.1120	0.7563	0.0863
UCF [17]	0.9034	0.0691	0.8876	0.0612	0.8181	0.1108	0.7296	0.1203
OURS	0.9065	0.0646	0.8993	0.0513	0.8214	0.1061	0.7574	0.0754

generated saliency maps are very close to the ground truth. This indicates that our PR curve obtains the higher precision than other algorithms. Moreover, the proposed algorithm achieves competitive results on both two challenging datasets: PASCAL-S and HKU-IS, which contain multiple salient objects in most of images. These advantages demonstrate that our saliency model is capable of predicting the salient objects accurately.

F-measure and MAE. We also report the results of quantitative comparison, as shown in Table 1. On the four datasets, our algorithm significantly outperforms most of other state-of-art algorithms over the evaluation metrics, especially in terms of MAE. For example, On ECSSD dataset, our method achieves the best performance. Comparison with the second best algorithm DCL, the MAE of our algorithm is conspicuously improved by 12.4% on DUT-OMRON dataset. This indicates that our algorithm is effective and robust to detect the salient object with complex backgrounds.

Visual results. We also provide a visual comparison of different algorithms, as shown in Fig. 4. It is obvious that our model shows superior capability of highlighting the salient objects and preserving the boundaries. In some challenging cases, our algorithm still produces favorable results with fewer noisy background, such as object near the image boundary (the second row), scattered backgrounds (the last row) and similar appearances between foreground and background (the third row). Besides, our model can generate more accurate saliency map with boundary preserved (the first row).

Table 2. Analysis of our saliency model.

Model Settings	ECSSD		DUT-OMRON	
	maxF	MAE	maxF	MAE
without CSM	0.8975	0.0651	0.7303	0.0824
with CSM	0.9065	0.0646	0.7570	0.0756

4.5. Model Analysis

To demonstrate the effectiveness of our deep model, we also evaluate the results of our algorithm with different model settings on ECSSD and DUT-OMRON datasets. The results are shown in Table 2. It can be seen that the model with CSM significantly improves the performance of saliency detection. Comparison with the model without CSM, our full model improves the maximum F-measure by 1% and 3.7% on ECSSD and DUT-OMRON datasets respectively, and simultaneously decreases the MAE by 1% and 8%.

5. CONCLUSIONS

In this paper, we propose a novel end-to-end deep saliency algorithm by exploiting the multi-path features, which lead to the performance improvement in saliency detection. Moreover, a coupled structure module is designed to obtain more fine-grained saliency results. The proposed model is efficient and can generate the accurate saliency maps without post-processing. Comprehensive experiments demonstrate the advantages of our MPFFNet. However, our model also has a limitation, that is some parts of the detected salient object is incomplete in challenging scenarios. In the future, we will consider to use the relationship between the object parts to improve the saliency detection.

6. REFERENCES

[1] Michael Donoser, Martin Urschler, Martin Hirzer, and Horst Bischof, "Saliency driven total variation segmentation," in *ICCV*, 2010, pp. 817–824.

[2] Vijay Mahadevan and Nuno Vasconcelos, "Biologically inspired object tracking using center-surround saliency mechanisms," *IEEE TPAMI*, vol. 35, no. 3, pp. 541–554, 2013.

[3] Shai Avidan and Ariel Shamir, "Seam carving for content-aware image resizing," *Acm Transactions on Graphics*, vol. 26, no. 3, pp. 10, 2012.

[4] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun, "Saliency optimization from robust background detection," in *CVPR*, 2014, pp. 2814–2821.

[5] Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, and Radomir Mech, "Minimum barrier salient object detection at 80 fps," in *ICCV*, 2015, p-p. 1404–1412.

[6] Huaizu Jiang, Jingdong Wang, Zejian Yuan, and Yang Wu, "Salient object detection: A discriminative regional feature integration approach," in *CVPR*, 2013, pp. 2083–2090.

[7] Hengliang Zhu, Bin Sheng, Xiao Lin, Yangyang Hao, and Lizhuang Ma, "Foreground object sensing for saliency detection," in *ACM ICMR*, 2016, pp. 111–118.

[8] Ali Borji, Dicky N Sihite, and Laurent Itti, "Salient object detection: a benchmark," *IEEE TIP*, vol. 24, no. 12, pp. 5706–5722, 2015.

[9] Guanbin Li and Yizhou Yu, "Visual saliency based on multiscale deep features," in *CVPR*, 2015, pp. 5455–5463.

[10] Gayoung Lee, Yu Wing Tai, and Junmo Kim, "Deep saliency with encoded low level distance map and high level features," in *CVPR*, 2016, pp. 660–668.

[11] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang, "Saliency detection by multi-context deep learning," in *CVPR*, 2015, pp. 1265–1274.

[12] Nian Liu and Junwei Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *CVPR*, 2016, pp. 678–686.

[13] Guanbin Li and Yizhou Yu, "Deep contrast learning for salient object detection," in *CVPR*, 2016, pp. 478–487.

[14] Kilian Q. Weinberger Laurens van der Maaten Gao Huang, Zhuang Liu, "Densely connected convolutional networks," in *CVPR*, 2017.

[15] L. Wang, H. Lu, X. Ruan, and M. H. Yang, "Deep networks for saliency detection via local estimation and global search," in *CVPR*, 2015, pp. 3183–3192.

[16] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Ruan Xiang, "Saliency detection with recurrent fully convolutional networks," in *ECCV*, 2016, pp. 825–841.

[17] Huchuan Lu Hongyu Wang Baocai Yin Pingping Zhang, Dong Wang, "Learning uncertain convolutional features for accurate saliency detection," in *ICCV*, 2017.

[18] X. Li, L. Zhao, L. Wei, M. H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "Deepsaliency: Multi-task deep neural network model for salient object detection.," *IEEE TIP*, vol. 25, no. 8, pp. 3919, 2016.

[19] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[20] Matthew D. Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, 2014, pp. 818–833.

[21] Fisher Yu and Vladlen Koltun, "Multi-scale context aggregation by dilated convolutions," in *ICLR*, 2016.

[22] Ming Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi Min Hu, "Global contrast based salient region detection," in *CVPR*, 2011, pp. 409–416.

[23] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia, "Hierarchical saliency detection," in *CVPR*, 2013, pp. 1155–1162.

[24] Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille, "The secrets of salient object segmentation," in *CVPR*, 2014, pp. 280–287.

[25] Chuan Yang, Lihe Zhang, Huchuan Lu, Ruan Xiang, and Ming Hsuan Yang, "Saliency detection via graph-based manifold ranking," in *CVPR*, 2013, pp. 3166–3173.

[26] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe:convolutional architecture for fast feature embedding," in *ACM MM*, 2014, pp. 675–678.