Fixation Guided Network for Salient Object Detection

Zhe Cui
cuizhe18@mails.ucas.ac.cn
University of Chinese Academy of Sciences

Li Su
suli@ucas.ac.cn
University of Chinese Academy of Sciences

Weigang Zhang
wgzhang@hit.edu.cn
Harbin Institute of Technology, Weihai

Qingming Huang
qmhuang@ucas.ac.cn
University of Chinese Academy of Sciences
Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences

ABSTRACT

Convolutional neural network (CNN) based salient object detection (SOD) has achieved great development in recent years. However, in some challenging cases, i.e. small-scale salient object, low contrast salient object and cluttered background, existing salient object detect methods are still not satisfying. In order to accurately detect salient objects, SOD networks need to fix the position of most salient part. Fixation prediction (FP) focuses on the most visual attractive regions, so we think it could assist in locating salient objects. As far as we know, there are few methods jointly consider SOD and FP tasks. In this paper, we propose a fixation guided salient object detection network (FGNet) to leverage the correlation between SOD and FP. FGNet consists of two branches to deal with fixation prediction and salient object detection respectively. Further, an effective feature cooperation module (FCM) is proposed to fuse complementary information between the two branches. Extensive experiments on four popular datasets and comparisons with twelve state-of-the-art methods show that the proposed FGNet well captures the main context of images and locates salient objects more accurately.

CCS CONCEPTS

• **Computing methodologies** → **Interest point and salient region detections**.

KEYWORDS

salient object detection, fixation prediction, convolutional neural network, computer vision

ACM Reference Format:
Zhe Cui, Li Su, Weigang Zhang, and Qingming Huang. 2021. Fixation Guided Network for Salient Object Detection. In *ACM Multimedia Asia (MMAsia '20), March 7–9, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3444685.3446288

MMAsia '20, March 7–9, 2021, Virtual Event, Singapore

ACM ISBN 978-1-4503-8308-0/21/03…$15.00
https://doi.org/10.1145/3444685.3446288

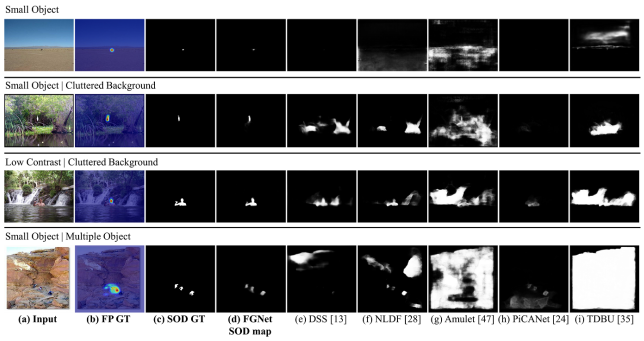

Figure 1: Examples to show some problems in current methods. (a) Input. (b) FP ground-truth. (c) SOD ground-truth. (d) SOD maps generated by the proposed FGNet. (e)-(i) SOD maps generated by DSS, NLDF, Amulet, PiCANet and TDBU respectively.

1 INTRODUCTION

Visual attention mechanism in human visual system means that people can focus the most attractive regions when looking at an image. Through mimicking the visual attention mechanism, salient object detection aims to segment the most visual distinctive objects in an image. As a fundamental task in computer vision, salient object detection is widely applied to many other visual tasks, such as image retrieval [11], semantic segmentation [38, 39], and image captioning [9].

During the past decades, a number of traditional methods [3, 43] were proposed to deal with SOD, which only use low-level cues and hand-crafted features to detect and segment salient objects. With the success of deep learning in computer vision, a number of CNN-based methods have been applied to salient object detection in recent years and have significantly outperformed traditional methods. Early CNN-based methods [19, 20, 31, 50] extract deep features and predict saliency scores for each image regions one by one, which is time-consuming. Currently, the most popular salient object detection methods [5, 10, 13, 22, 24, 28, 33–37, 40, 41, 46–48] are based on the fully convolutional networks (FCN) [27]. These FCN-based salient object detection methods focus on exploring different feature fusion strategies and have achieved satisfactory performance.

Although significant progresses have been achieved, some salient objects can hardly be precisely discovered in challenging cases as shown in Fig. 1. This is because the visual contrast between salient

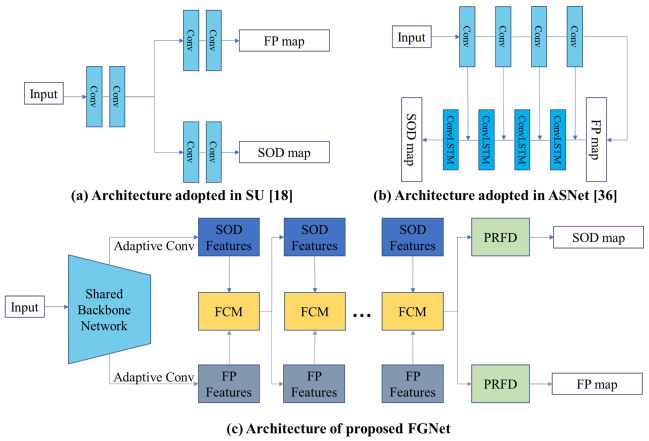

Figure 2: Architecture comparison between the proposed FGNet and previous models that combine SOD and FP. (a) Architecture adopted in SU only share features in low layers (b) Architecture adopted in ASNet use FP to help SOD unidirectional. (c). Architecture of the proposed FGNet can mutually exchange information between SOD and FP with cascade FCMs.

objects and background is not obvious. Under these circumstances, human visual attention is critical to discovering salient objects. In order to detect and segment salient objects accurately, we must to locate attention-grabbing regions firstly. FP aims to predict fixation points where humans look during scene free viewing, while SOD takes a further step to segment the whole extent of salient objects. The corresponding pixels with higher value in both FP and SOD maps are more likely to be attended. In addition, the analyses in [2, 4] indicate the intrinsic correlation between FP and SOD.

However, most saliency researches treat SOD and FP as two individual tasks and only few works [18, 36] try to explore the relationship between them. Kruthiventi et al. [18] implemented both fixation prediction and salient object detection via a two-branch unified network SU which share features only in low layers as shown in Fig. 2 (a). Wang et al. [36] presented an attentive saliency network ASNet to progressively refine saliency map from fixation map through a top-down pathway by aggregating multi-level features as shown in Fig. 2 (b).

Based on the above observation, we focus on leveraging the complementarity between salient object detection and fixation prediction information. We propose a fixation guided salient object detection network (FGNet) to couple salient object detection and fixation prediction in a unified network as shown in Fig. 2 (c). Compared with SU [18] and ASNet [36], the proposed FGNet utilize advantages of both fixation prediction and salient object detection by building a bidirectional information flow architecture. Instead of only using fixation features to guide salient object detection, we optimize both features by mutually flowing information between SOD and FP.

In summary, the contributions are as follows:

- We propose FGNet to explicitly combine complementary information between salient object detection and fixation prediction. Compared with exiting methods only using FP as an auxiliary task to unidirectionally improve the representation ability of SOD features, we propose cascade feature cooperation modules to exchange information between SOD and FP bidirectional.
- The proposed FGNet promotes SOD and FP at the same time by allowing them to mutually pass information to each other, yielding more accurate SOD maps.
- Experimental results on four popular datasets show that the proposed FGNet substantially improves the SOD performance compared with 12 state-of-the-art methods.

2 RELATED WORK

FP is an active research topic in computer vision area for a long time. Itti et al. [16] proposed a bottom-up attention models based on psychological theories and started the field of visual attention. Traditional fixation prediction models [45] were commonly built on bottom-up structure using stimulus-driven biologically features and certain heuristics. Recent deep learning based fixation prediction models [15, 23] leverage CNN to extract representative features for fixation prediction by aggregating features from multi-stream, multi-scale and multi-level, etc.

Compared with fixation prediction, the history of SOD is short and the original works of SOD can trace back to the work of Liu et al. [26] and Achanta et al. [1]. Traditional salient object detection methods rely on hand-crafted features to predict saliency scores, such as center prior [17], contrast prior [6, 30], and spectral information [14]. These approaches can usually extract hand-crafted features with little time cost. However, the hand-crafted features are low-level and hardly capture high-level semantic knowledge of the salient objects, which will decrease detection accuracy.

With the popular of the CNN which has powerful feature extraction capability, SOD began to leverage CNN to extract features and got great progress. Zhao et al. [50] presented a multi-context deep learning framework to extract local and global context simultaneously, which were then fed into CNN for saliency classification. Wang et al. [31] used local estimation and global contrast information to produce saliency maps. These methods fed each processing unit into classifiers for saliency score prediction and have obvious disadvantages: time-consuming, unable to use overall spatial information, and all pixels in each processing unit share the same saliency score.

Inspired by the great success of FCN in semantic segmentation, salient object detection area has turned attention to FCN. FCN can extract multi-level features, where high-level features capture semantic knowledge and low-level features contain more detailed information. Researchers focus on designing different fusion strategies to aggregate multi-level features. Wang et al. [33] used salient priors to make the training of network easier, utilized cascaded FCNs to refine saliency map iteratively by correcting its previous errors, until the final prediction was generated in the last time step. Hou et al. [13] proposed a new method to fuse the low-level features and the high-level features by adding several short connections from deeper side-outputs to shallower ones based on the Holistically-Nested Edge Detector [42]. Chen et al. [5] proposed a reverse attention network to compensate the missing parts between the prediction and the ground-truth by erasing current generated

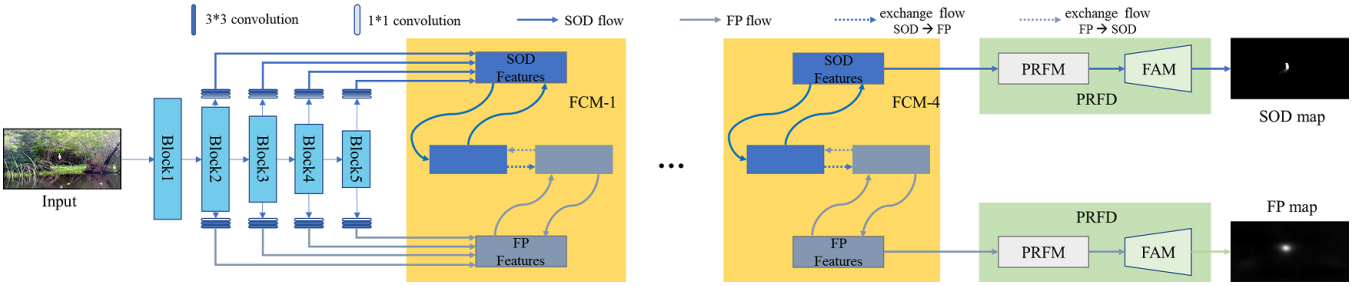

Figure 3: The overall architecture of FGNet. There are two branches (SOD branch and FP branch) based on the shared backbone network. FGNet consists of four cascade Feature Cooperation Modules (FCM) and two Pyramid Receptive Field Decoders (PRFD). FCM is proposed to exchange information between two branches. PRFD is used to generate SOD and FP maps by fusing refined features.

saliency maps. Most of aggregation-based methods fuse all features extracted from FCN to generate final saliency map. Wu et al. [40] discovered that fusing features in shallow layers brings little improvement in the final saliency map, but increases computation cost greatly.

3 FGNET

In this section, we first introduce the overall architecture of the proposed network. Then, we show details about how to fuse the features between two tasks and how to integrate the refined features to generate SOD and FP maps.

3.1 Overall Architecture

We choose ResNet-50 [12] as a common feature extractor and modify it to meet the SOD requirements. ResNet-50 consists of 49 convolutional layers with five convolutional blocks, following a global pooling layer and a fully connected layer. We only use five residual blocks to extract multi-scale features and denote them as $R = \{R_i | i = 1, 2, 3, 4, 5\}$. The block size is $\frac{W}{2^i} \times \frac{H}{2^i} \times C_i$, where H, W are the width and height of the input image, and C_i is the channel number of the i-th feature R_i. Since shallow layer contribute less to the final results but largely increase the computation cost as demonstrated in [40], so only $\{R_i | i = 2, 3, 4, 5\}$ are reserved. In addition, we add three convolutional layers on the top of each block before the branch of fixation prediction and salient object detection separately. On the one hand, it is for the sake of making the features extracted from the shared backbone network more adaptive to the two tasks. On the other hand, it is for changing the channel to 32 in order to reduce computation cost. Now we get two feature groups $S = \{S_i | i = 2, 3, 4, 5\}$ and $F = \{F_i | i = 2, 3, 4, 5\}$ for SOD and FP respectively.

FGNet is a two-branch architecture network: salient object detection branch and fixation prediction branch, as shown in Fig. 3. Salient object detection branch focuses on extracting fine-grained visual features to precisely segment salient objects. Fixation prediction branch aims to capture the global image information to locate salient objects. Since both SOD and FP are related to human attention, features from these two branches have inherent relevance and complementarity. We further propose a feature cooperation module (FCM) to facilitate information sharing and exchanging between the two branches. After the features have learned enough, we feed them into the pyramid receptive field decoder (PRFD) to generate SOD and FP maps.

3.2 Feature Cooperation Module

In order to make full use of the complementarity between SOD and FP features, the feature cooperation module (FCM) is proposed. FCM consists of two sub-modules: mutual feature exchange module (MFEM) and internal feature fusion module (IFFM) as shown in Fig. 4. MFEM builds a bidirectional information flow mechanism aiming to exchange information between features of two branches in each level. Considering that low-level features contain more background distractors, we only choose the equivalent and higher level features when aggregating complementary features to suppress distractors. In addition, we observe that there are some fixation points fall outside salient objects occasionally, so the FP features need to multiplied by corresponding SOD features before aggregated into SOD features. IFFM offers a top-down feature fusion mechanism within a single branch. IFFM further explicitly increases the weight of high-level information in each level features, to make sure each layer can capture enough location information by introducing high-level guiding flows between adjacent layers.

The detailed illustration of FCM is shown in Fig. 4. FCM takes both SOD features S and FP features F as input. After the process of MFEM and IFFM, FCM generates refined SOD features $S^{'}$ and FP features $F^{'}$. By stacking multiple FCMs (i.e., the output of one FCM is used as the input to the next FCM), both SOD and FP can learn more complementary information from each other. To achieve the best trade-off between performance and model size, we stack four FCMs.

3.3 Pyramid Receptive Field Decoder

By stacking multiple FCMs, complementary features in two branches have been conducted a full exchange of information. Another question deserves considering is how to fuse these multi-level refined features.

As demonstrated in [49], the receptive fields of CNN layers are much smaller compared with its theoretical value especially when CNNs go deeper. In order to solve this problem and allow each pixel in feature maps could capture different receptive fields of feature maps, we propose a pyramid receptive field module (PRFM) inspired by [25]. As shown in Fig. 5, PRFM comprises four parallel branches containing convolutional layers of different kernel size

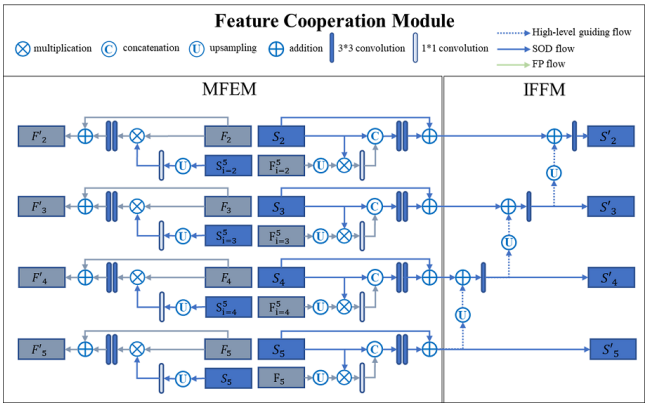

Figure 4: Detailed illustration of FCM. It comprises mutual feature exchange module (MFEM) and internal feature fusion module (IFFM). MFEM is developed to fuse complementary features from SOD and FP. IFFM is designed to further enhance the high-level location information.

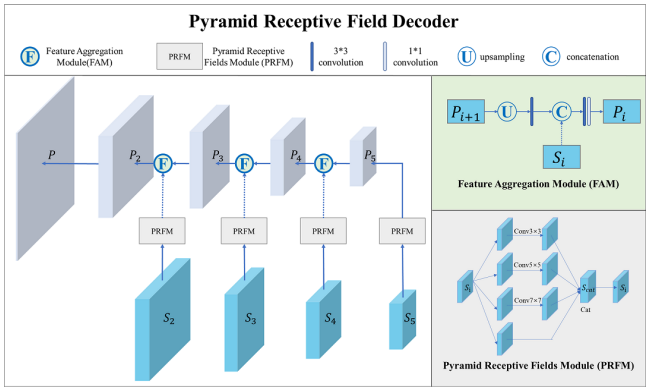

Figure 5: Detailed illustration of PRFD. It comprises pyramid receptive field module (PRFM) and feature aggregation module (FAM). PRFM is used to capture different receptive fields of feature maps. FAM is designed to seamlessly aggregate multi-scale features.

$\{1, 3, 5, 7\}$ to capture pyramid receptive fields for each pixel in feature maps. Then we utilize feature aggregation module (FAM) to progressively aggregate high-level features with low-level features in a top-down pathway. At each time step, the fused feature map P_i can be generated by combining S_i and feature maps $P_{(i+1)}$ from previous step using FAM. The feature maps P_2 in the last step has $[\frac{W}{4}, \frac{H}{4}]$ size and 32 channels, so additional convolutional layers and upsample layers are added to generate final SOD map P. The generation process of FP map Q is similar to that of SOD map.

After get SOD map P and FP map Q, the total train loss of FGNet L_{total} could be calculated by adding salient object detection loss $L_{CE}(P, GT_s)$ and fixation prediction loss $L_{CE}(Q, GT_f)$ as (1):

$$L_{total} = L_{CE}(P, GT_s) + L_{CE}(Q, GT_f) \tag{1}$$

where $GT = \{GT_s, GT_f\}$ are the ground-truth map for SOD and FP, $\theta = \{\theta_s, \theta_f\}$ are the parameters corresponding to maps $\{P, Q\}$, and L_{CE} is the standard pixel-wise cross entropy loss formulated as (2):

$$\begin{aligned} L_{CE}(P, GT_s|\theta_s) &= -\sum_{i=1}^{N} GT_s^i log(P^i) + (1 - GT_s^i) log(1 - P^i) \\ L_{CE}(Q, GT_f|\theta_f) &= -\sum_{i=1}^{N} GT_f^i log(Q^i) + (1 - GT_f^i) log(1 - Q^i) \end{aligned} \tag{2}$$

where N is the pixel number.

4 EXPERIMENTS

4.1 Datasets and Evaluation Metric

To train and evaluate the proposed FGNet, five popular benchmark datasets are adopted, including ECSSD [43], HKU-IS [20], PASCAL-S [21], DUTS [32], DUT-OMRON [44]. **ECSSD** contains 1000 semantically meaningful and structure complex images. **HKU-IS** contains 4447 images with high quality annotations, which have multiple disconnected salient objects or objects touching image boundary. **PASCAL-S** contains 850 images selected from PASCAL VOC 2010. **DUTS** contains 10553 images for training and 5019 images for testing, and it is the largest SOD benchmark dataset. **DUT-OMRON** contains 5168 high-quality images which have annotations for both salient object detection and fixation prediction.

Table 1: Ablation analysis on DUTS-TEST dataset. Best scores in each column are highlighted in bold. FCM-4 is the adopted architecture

variant	maxF	meanF	MAE	S
single SOD branch	0.863	0.765	0.052	0.861
FCM-1	0.885	0.795	0.043	0.879
FCM-2	0.889	0.801	0.041	0.887
FCM-4	0.892	**0.815**	**0.038**	**0.891**
FCM-6	**0.895**	0.812	0.039	0.889
w/o PRFM	0.884	0.799	0.041	0.883

Considering that we focus on the task of SOD, so we just give visual results for FP to verify the effectiveness of the generated FP maps. And for SOD task, we adopt four widely used evaluation metrics including mean absolute error (MAE), max F-measure (max F), mean F-measure (mean F) and precision-recall curve in order to compare with other methods. Further, we also evaluate the proposed model based on the S-measure [7], E-measure [8] and weighted F-measure [29] to give more comprehensive evaluation.

F-measure is a balanced mean of average precision and average recall which can be calculated by (3), where β^2 is set to 0.3 to weigh precision more than recall as suggested in [1].

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

4.2 Implementation Details

Firstly, we adopt DUT-OMRON dataset to pre-train FGNet because this dataset could provide both SOD and FP ground-truths. Then, FGNet is trained on DUTS-TRAIN dataset following most salient object detection works [24, 34, 40, 41, 48]. Since DUTS-TRAIN only contains SOD ground-truth, so the total train loss of equals to salient object detection loss $L_{total} = L_{CE}(P, GT_s)$ in this stage.

Table 2: Comparison with state-of-the-art methods. Max F-measure (maxF, larger is better), Mean F-measure (meanF, larger is better), MAE (smaller is better), S measure (larger is better) are used to measure the model performance. '-' denotes that the authors have not provided corresponding saliency maps. * means methods combine FP and SOD. The top three results are marked in red, blue, and green, respectively. FGNet achieves the state-of-the-art under all evaluation metrics on four popular datasets.

method	ECSSD				HKU-IS				PASCAL-S				DUTS-TEST			
	maxF	meanF	MAE	S	maxF	meanF	MAE	S	maxF	meanF	MAE	S	maxF	meanF	MAE	S
MDF	0.832	0.807	0.105	0.776	0.860	0.784	0.129	0.810	0.764	0.705	0.145	0.688	-	-	-	-
RFCN	0.890	0.834	0.107	0.852	0.893	0.835	0.089	0.859	0.829	0.747	0.132	0.794	0.784	0.711	0.090	0.791
DSS	0.908	0.865	0.062	0.883	0.898	0.854	0.051	0.879	0.824	0.763	0.103	0.797	0.813	0.712	0.065	0.826
NLDF	0.905	0.874	0.063	0.875	0.900	0.872	0.049	0.876	0.826	0.770	0.099	0.798	0.813	0.738	0.065	0.816
Amulet	0.915	0.870	0.059	0.894	0.894	0.838	0.053	0.882	0.832	0.764	0.097	0.815	0.779	0.672	0.085	0.803
BMPM	0.928	0.894	0.044	0.911	0.920	0.875	0.039	0.906	0.857	0.803	0.073	0.840	0.852	0.762	0.049	0.861
PAGR	0.927	0.894	0.061	0.889	0.918	0.886	0.048	0.887	0.851	0.803	0.092	0.813	0.854	0.783	0.056	0.838
DGRL	0.925	0.903	0.043	0.906	0.913	0.882	0.037	0.897	0.853	0.807	0.074	0.834	0.828	0.794	0.050	0.842
PiCANet-R	0.935	0.886	0.046	0.917	0.918	0.870	0.043	0.904	0.863	0.798	0.075	0.849	0.860	0.759	0.051	0.869
PiCANet	0.931	0.885	0.046	0.914	0.921	0.870	0.042	0.906	0.862	0.796	0.076	0.845	0.851	0.749	0.054	0.861
PAGE	0.931	0.906	0.042	0.912	0.918	0.882	0.037	0.903	0.852	0.810	0.077	0.835	0.838	0.777	0.052	0.854
TDBU	0.938	0.88	0.041	0.918	0.922	0.878	0.038	0.907	0.859	0.779	0.071	0.844	0.855	0.767	0.048	0.865
ASNet*	0.932	0.875	0.047	0.915	0.922	0.872	0.041	0.906	0.871	0.791	0.069	0.856	0.835	0.728	0.061	0.843
FGNet*	0.948	0.916	0.037	0.926	0.936	0.898	0.033	0.917	0.879	0.833	0.064	0.861	0.892	0.815	0.038	0.891

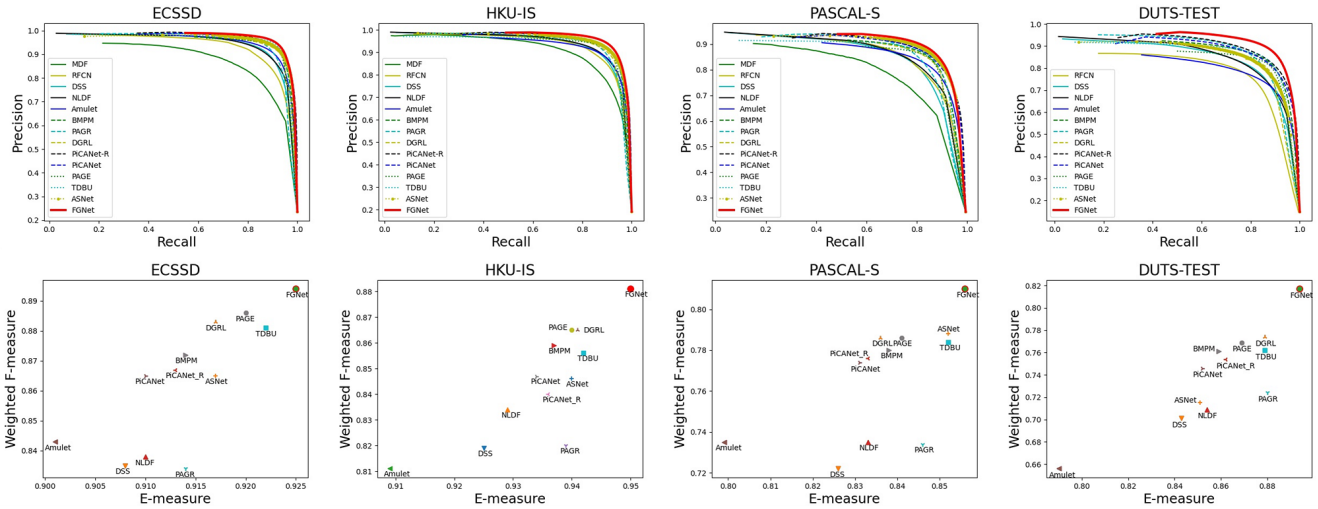

Figure 6: Performance comparison with state-of-the-art methods on four popular benchmark datasets. The first row shows precision-recall curves. The second row shows weighted F-measure and E-measure. It can be seen that the proposed method performs favorably against state-of-the-arts.

Remaining datasets DUTS-TEST, ECSSD, PASCAL-S, HKU-IS are used to evaluate the proposed model.

The parameters of the backbone are initialized by ResNet-50 pre-trained on ImageNet. For all newly added convolution layers, their weights are initialized by normal distribution with standard deviation = 0.01 and mean value = 0. The proposed model is implemented in PyTorch. The whole network is trained by stochastic gradient descent (SGD). Momentum and weight decay are set to 0.9 and 5e-4. Epoch and batch size are set to 30 and 8 for both pretrain and train stages. In the pretrain stage, learning rate is initialized as 2e-3 and decreased by 10% at 20 epochs, and in the train stage, learning rate is initialized as 2e-4 and decreased by 10% at 20 epochs.

4.3 Ablation Study

In this subsection, We conduct a series of ablation experiments to confirm the effectiveness of our proposed FGNet. As shown in Tab. 1, we analyzing the contribution of each part on DUTS-TEST. Firstly, we evaluate the performance of the proposed model with only SOD branch to show the advantage of fusing two branches. Secondly, we explore how the number of FCMs influence the final detection results. Finally, we analyse how about the performance would be when replacing the PRFD with simple fusion strategies. It demonstrates that all modules can help locate and segment salient objects.

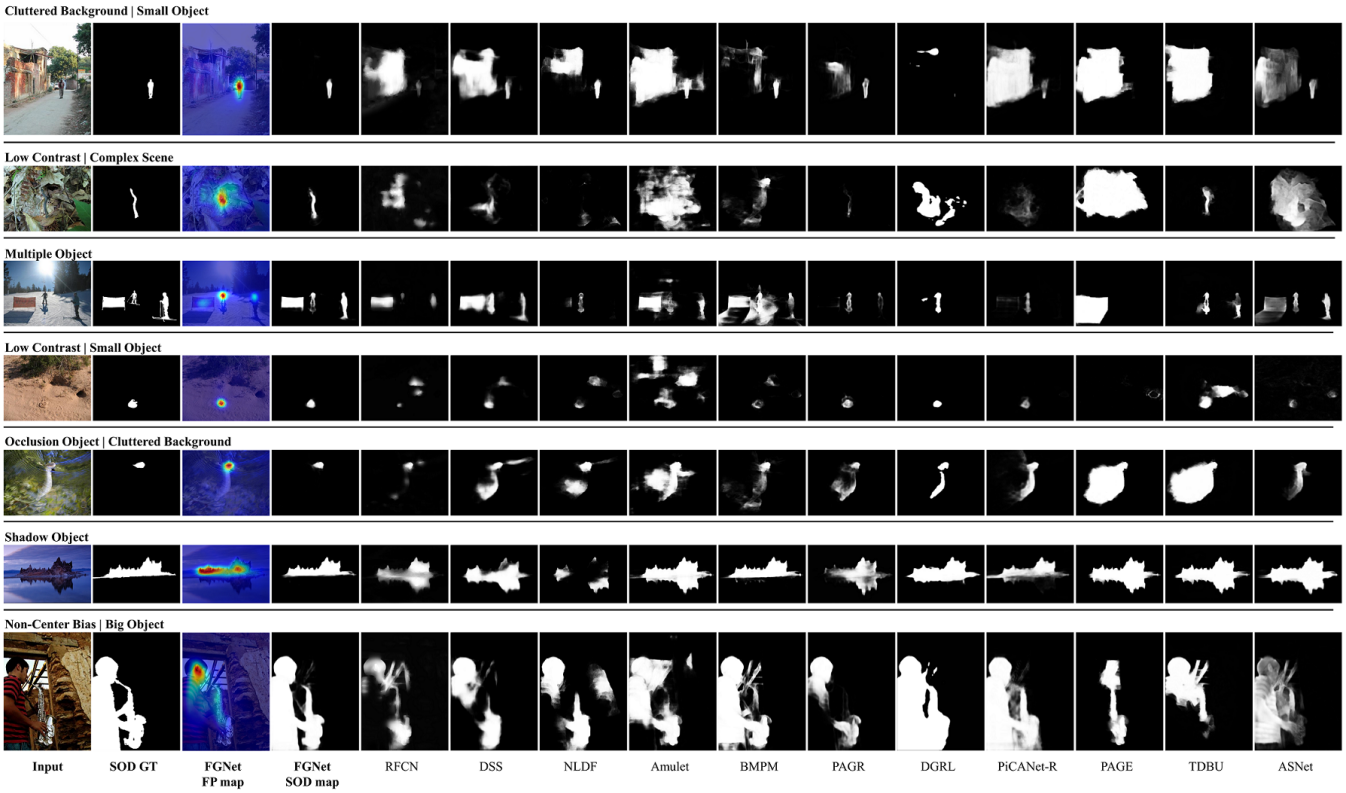

Figure 7: Visual comparison of FGNet results compared with other state-of-the-art methods. Each row represents one image and we highlight the main challenges. Each column represents the results of one method, where ASNet and the proposed FGNet combine FP and SOD. (GT: ground-truth)

4.4 Comparison with the State-of-the-arts

Quantitative Comparison: We compare the proposed method with 12 state-of-the-art salient object detection methods including MDF [20], RFCN [33], DSS [13], NLDF [28], Amulet [47], BMPM [46], PAGR [48], DGRL [34], PiCANet [24], PAGE [37], TDBU [35], ASNet [36]. Among these methods, ASNet and the proposed FGNet are methods that combine FP and SOD. All the saliency maps are provided by authors or generated by running their pre-trained models with parameter recommended in their papers. For fair comparison, we evaluate all the saliency maps with the same evaluation codes. Tab. 2 shows the max F-measure, mean F-measure, MAE, and S-measure. Fig. 6 shows the precision-recall curves in the first row, weighted F-measure and E-measure in the second row. FGNet achieves state-of-the-art performance on four datasets without any post-processing techniques.

Visual Comparison: As showed in Fig. 7, we can see that FGNet performs better when dealing with various challenging cases. For the first, second and fourth samples, it is difficult to detect salient objects only based on low-level visual features due to cluttered background, small salient objects and low contrast between foreground and background. For the fifth and sixth samples, other methods mistakenly detect hidden parts underwater and shadows as salient objects. Because they and the salient objects can be seen as a whole in terms of the low-level visual characteristics. However, benefiting from the complementary fixation information which mimics human visual mechanisms, FGNet can detect the most important object(s) accurately. It is worth mentioning that thanks to the cooperation of FP and SOD features, FGNet could not only locate salient region but also segment the whole salient object clearly.

5 CONCLUSION

In this paper, we propose a fixation guided salient object detection network FGNet to leverage the correlation between SOD and FP. Different from other methods which just take fixation map as an auxiliary information, the complementarity between SOD and FP is fully considered. Firstly, the SOD and FP features are extracted by the two-branch network. Then the FCMs are proposed to fuse the complementary information by building a mutual information exchange mechanism. Finally, the PRFDs are employed to integrate these refined features and generate SOD and FP maps. Extensive experiments demonstrate that FGNet outperforms most of the state-of-the-art approaches on four popular datasets.

ACKNOWLEDGMENTS

This work was supported in part by the National Key RD Program of China under Grant 2018AAA0102003 and 2018YFE0118400, in part by National Natural Science Foundation of China: 61931008, 61976069, and 61771457, and in part by the Fundamental Research Funds for Central Universities.

REFERENCES

[1] Radhakrishna Achanta, Sheila S. Hemami, Francisco J. Estrada, and Sabine Süsstrunk. 2009. Frequency-tuned salient region detection. In *CVPR*. 1597–1604.

[2] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. 2015. Salient Object Detection: A Benchmark. *IEEE Transactions on Image Processing* 24, 12 (2015), 5706–5722.

[3] Ali Borji and Laurent Itti. 2012. Exploiting local and global patch rarities for saliency detection. In *CVPR*. 478–485.

[4] Ali Borji, Dicky N.Sihite, and Laurent Itti. 2013. What stands out in a scene? A study of human explicit saliency judgment. *Vision Research* 91, 15 (2013), 62–77.

[5] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. 2018. Reverse Attention for Salient Object Detection. In *ECCV (9) (Lecture Notes in Computer Science)*, Vol. 11213. 236–252.

[6] Ming-Ming Cheng, Guo-Xin Zhang, Niloy J. Mitra, Xiaolei Huang, and Shi-Min Hu. 2011. Global contrast based salient region detection. In *CVPR*. 409–416.

[7] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. 2017. Structure-measure: A new way to evaluate foreground maps. In *ICCV*. 4558–4567.

[8] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. 2018. Enhanced-alignment measure for binary foreground map evaluation. In *IJCAI*. 698–704.

[9] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2015. From captions to visual concepts and back. In *CVPR*. 1473–1482.

[10] Mengyang Feng, Huchuan Lu, and Errui Ding. 2019. Attentive Feedback Network for Boundary-Aware Salient Object Detection. In *CVPR*. 1623–1632.

[11] Junfeng He, Jinyuan Feng, Xianglong Liu, Tao Cheng, Tai-Hsu Lin, Hyunjin Chung, and Shih-Fu Chang. 2012. Mobile product search with Bag of Hash Bits and boundary reranking. In *CVPR*. 3005–3012.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, , and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. 770–778.

[13] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip H. S. Torr. 2019. Deeply Supervised Salient Object Detection with Short Connections. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 4 (2019), 815–828.

[14] Xiaodi Hou and Liqing Zhang. 2007. Saliency detection: A spectral residual approach. In *CVPR*. 1–8.

[15] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. 2015. SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *CVPR*. 262–270.

[16] Laurent Itti, Christof Koch, and Ernst Niebur. 1998. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 20, 11 (1998), 1254–1259.

[17] Zhuolin Jiang and Larry S. Davis. 2013. Submodular salient region detection. In *CVPR*. 2043–2050.

[18] Srinivas S. S. Kruthiventi, Vennela Gudisa, Jaley H. Dholakiya, and R. Venkatesh Babu. 2016. Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation. In *CVPR*. 5781–5790.

[19] Gayoung Lee, Yu-Wing Tai, and Junmo Kim. 2016. Deep Saliency with Encoded Low level Distance Map and High Level Features. In *CVPR*. 660–668.

[20] Guanbin Li and Yizhou Yu. 2015. Visual saliency based on multiscale deep features. In *CVPR*. 5455–5463.

[21] Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille. 2014. The Secrets of Salient Object Segmentation. In *CVPR*. 280–287.

[22] Jiangjiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. 2019. A Simple Pooling-Based Design for Real-Time Salient Object Detection. In *CVPR*. 3917–3926.

[23] Nian Liu, Junwei Han, Tianming Liu, and Xuelong Li. 2018. Learning to predict eye fixations via multiresolution convolutional neural networks. *IEEE Transactions on Neural Networks and Learning Systems* 29, 2 (2018), 392–404.

[24] Nian Liu, Junwei Han, and Ming-Hsuan Yang. 2018. PiCANet: Learning PixelWise Contextual Attention for Saliency Detection. In *CVPR*. 3089–3098.

[25] Songtao Liu, Di Huang, and Yunhong Wang. 2018. Receptive field block net for accurate and fast object detection. In *ECCV*.

[26] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. 2007. Learning to detect a salient object. In *CVPR*. 1–8.

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*. 3431–3440.

[28] Zhiming Luo, Akshaya Mishra, Andrew Achkar, Justin Eichel, Shaozi Li, and Pierre-Marc Jodoin. 2017. Non-Local Deep Features for Salient Object Detection. In *CVPR*. 6593–6601.

[29] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. 2014. How to evaluate foreground maps. In *CVPR*. 248–255.

[30] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. 2012. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*. 733–740.

[31] Lijun Wang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. 2015. Deep Networks for Saliency Detection via Local Estimation and Global Search. In *CVPR*. 3183–3192.

[32] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. 2017. Learning to Detect Salient Objects with Image-Level Supervision. In *CVPR*. 3796–3805.

[33] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Xiang Ruan. 2016. Saliency Detection with Recurrent Fully Convolutional Networks. In *ECCV (4) (Lecture Notes in Computer Science)*, Vol. 9908. 825–841.

[34] Tiantian Wang, Lihe Zhang, Shuo Wang, Huchuan Lu, Gang Yang, Xiang Ruan, and Ali Borji. 2018. Detect Globally, Refine Locally: A Novel Approach to Saliency Detection. In *CVPR*. 3127–3135.

[35] Wenguan Wang, Jianbing Shen, Ming-Ming Cheng, and Ling Shao. 2019. An iterative and cooperative top-down and bottom-up inference network for salient object detection. In *CVPR*. 5968–5977.

[36] Wenguan Wang, Jianbing Shen, Xingping Dong, Ali Borji, and Ruigang Yang. 2020. Inferring salient objects from human fixations. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 8 (2020), 1913–1927.

[37] Wenguan Wang, Shuyang Zhao, Jianbing Shen, Steven C. H. Hoi, and Ali Borji. 2019. Salient Object Detection With Pyramid Attention and Salient Edges. In *CVPR*. 1448–1457.

[38] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. 2018. Weakly-supervised semantic segmentation by iteratively mining common object features. In *CVPR*. 1354–1362.

[39] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. 2017. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *CVPR*. 6488–6496.

[40] Zhe Wu, Li Su, and Qingming Huang. 2019. Cascaded Partial Decoder for Fast and Accurate Salient Object Detection. In *CVPR*. 3902–3911.

[41] Zhe Wu, Li Su, and Qingming Huang. 2019. Stacked cross refinement network for edge-aware salient object detection. In *ICCV*. 7263–7272.

[42] Saining Xie and Zhuowen Tu. 2015. Holistically-nested edge detection. In *ICCV*. 1395–1403.

[43] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. 2013. Hierarchical Saliency Detection. In *CVPR*. 1155–1162.

[44] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. 2013. Saliency Detection via Graph-Based Manifold Ranking. In *CVPR*. 3166–3173.

[45] Jianming Zhang and Stan Sclaroff. 2013. Saliency Detection: A Boolean Map Approach. In *ICCV*. 153–160.

[46] Lu Zhang, Ju Dai, Huchuan Lu, You He, and Gang Wang. 2018. A Bi-Directional Message Passing Model for Salient Object Detection. In *CVPR*. 1741–1750.

[47] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. 2017. Amulet: Aggregating Multi-level Convolutional Features for Salient Object Detection. In *ICCV*. 202–211.

[48] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. 2018. Progressive Attention Guided Recurrent Network for Salient Object Detection. In *CVPR*. 714–722.

[49] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid scene parsing network. In *CVPR*. 6230–6239.

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