2021 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-4115-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIP42928.2021.9506100

MULTI-SCALE GRAPH CONVOLUTIONAL INTERACTION NETWORK FOR SALIENT OBJECT DETECTION

Wenqi Che *Luoyi Sun* *Zhifeng Xie*⋆ *Youdong Ding*⋆ *Kaili Han*

Shanghai University, Shanghai, China

ABSTRACT

Remarkable progress has been achieved for salient object detection based on deep learning. However, most of the previous works have the issues of how to extract more effective information from scale-varying data and how to improve the boundary quality. In this paper, we propose the multi-scale graph convolutional interaction network (MGCINet), which consists of the feature interaction module (FIM), the feature aggregation module (FAM), and the residual refinement module (RRM). FIMs fuse interactive features from neighboring scales. Based on two-layers graph convolutional network, FAMs aggregate scale-specific information by graph nodes interaction. RRMs optimize the coarse saliency maps with blurred boundaries by U-net residual blocks. In addition, we propose multi-scale weighted structural loss to assign different weights to pixels while focusing on image structure at various scales. Experiments show that our method outperforms the state-of-the-arts on five benchmark datasets under different evaluation metrics.

Index Terms— Salient object detection, graph convolutional network, multi-scale interaction

1. INTRODUCTION

Salient object detection (SOD) aims to distinguish the most visually obvious regions. It has been widely used in computer vision field with the development of deep learning, such as visual tracking [1], semantic segmentation [2], non-photorealistic rendering [3] and so on.

Benefiting from the powerful feature extraction capability of convolutional neural networks (CNN), traditional salient detection methods based on hand-crafted features [4, 5] are gradually being surpassed. Recently, most of the models have been implemented by fusing multi-scale features extracted by CNN. For example, Wang et al. [6] proposed a progressive feature polishing network that polishes multi-scale features in parallel by simple structures in multiple layers. Pang et al. [7] proposed a multi-scale interactive network with the transformation-interaction-fusion strategy to better extract multi-scale features. However, those previous works still face a key issue of how to extract and aggregate more effective information from scale-varying data. Besides, another issue is that most existing models focus more on region accuracy rather than boundary quality, which may fail to yield the clear and accurate boundary segmentation of salient objects.

In this paper, we propose multi-scale graph convolutional interaction network (MGCINet) for higher-quality salient object detection. For the first issue, we first utilize the feature interaction module to perform fusion of interactive features from neighboring scales to obtain initial feature maps. Then we further propose the feature aggregation module consisting of two-layers graph convolutional network (GCN) to aggregate scale-specific information by graph nodes interaction. The module can expand the receptive field and aggregate the features of the neighboring nodes through the message propagation mechanism. For the second issue, we design the residual refinement module based on U-net structure to optimize the coarse saliency maps with blurred boundaries by learning the residuals between the saliency predictions and the ground truth. In addition, the traditional binary cross entropy loss treats pixels equally and ignores the overall image structure in most SOD models. Thus, we reconstruct the multi-scale weighted structural loss to emphasize multi-scale image structure and assign distinct weights to pixels, which can guide the network to focus on relatively more local details.

Our contributions can be summarized as follows:

- We propose a novel MGCINet to more precisely extract multi-scale interactive features while performing higher-accuracy boundary segmentation.
- We construct the feature aggregation module based on two-layers GCN, which can effectively aggregate scale-specific features through graph nodes interaction.
- We design the residual refinement module with U-net residual blocks to refine fuzzy boundaries in the coarse saliency maps.
- We propose the multi-scale weighted structural loss to concentrate on local details by treating pixels unequally and emphasizing image structure at various scales.

2. METHOD

In this paper, we propose a multi-scale graph convolutional interaction network (MGCINet) for salient object detection,

⋆Corresponding authors: zhifeng_xie@shu.edu.cn, ydding@shu.edu.cn

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on May 11,2022 at 02:17:28 UTC from IEEE Xplore. Restrictions apply.

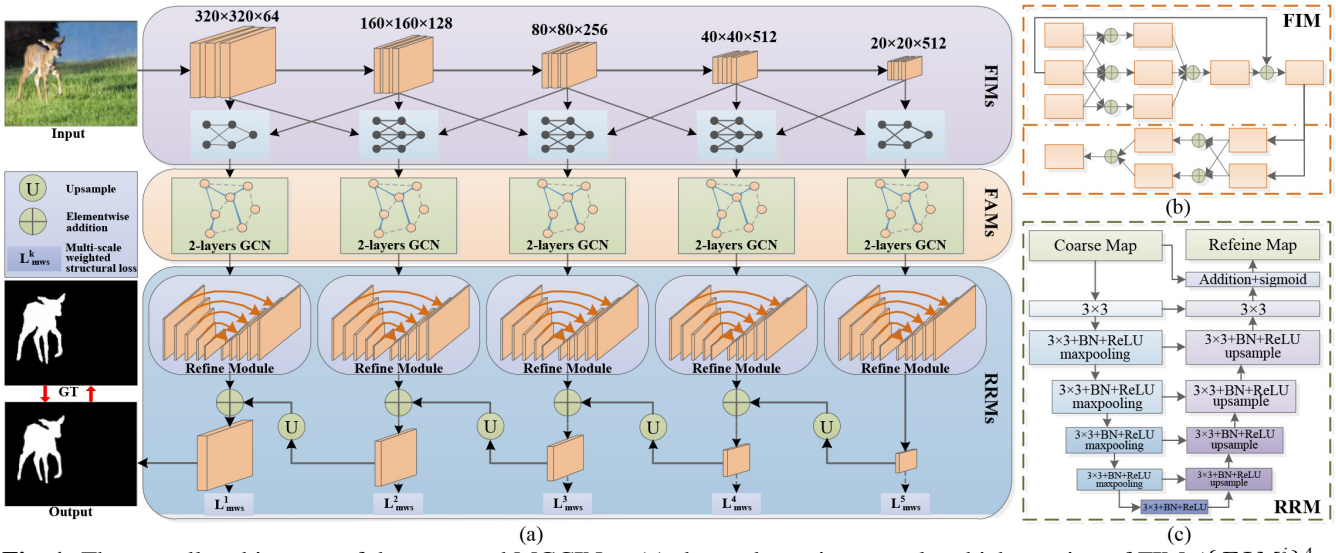

Fig. 1. The overall architecture of the proposed MGCINet. (a) shows the main network, which consists of FIMs($\{FIM^i\}_{i=0}^4$) for feature extraction, FAMs($\{FAM^i\}_{i=0}^4$) for feature aggregation, and RRMs($\{RRM^i\}_{i=0}^4$) for optimization. (b) shows the network structure of FIM and (c) shows the network structure of RRM.

and the overall network structure is shown in Fig.1(a). The network mainly consists of FIMs, FAMs and RRMs. Besides, we reconstruct the multi-scale weighted structural loss to supervise the training stage.

2.1. Feature Interaction Module

Inspired by MINet [7], the network structure of the feature interaction module (FIM) is shown in Fig.1(b). FIMs take features from neighboring scales as input and perform convergent of features interaction. This process can utilize the convolutional features at various scales to produce the initial feature maps. Then the feature maps are fed into FAMs for graph nodes interaction to obtain scale-specific information.

2.2. Feature Aggregation Module

The feature aggregation module (FAM) mainly consists of two-layers graph convolutional network (GCN), and the network structure is shown in Fig.2. FAMs performs feature aggregation at scale-specific with GCN by graph nodes interaction. GCN is a form of Laplace smoothing, which cannot be too deep to perform excessive smoothing, so we construct a two-layers GCN. Here, we require a graph model, i.e., the feature matrix of the nodes and the adjacency matrix that represents the relationship between the nodes. The image grid data obtained by FIMs needs to be expanded into graph structure data. The graph structure data can be represented as a binary group $G(V, E)$, V represents the set of nodes of the graph and is a $|N| * S$ feature matrix, $|N|$ is the number of nodes of the graph and S is the dimension of the node feature vector. E is the set of edges of the graph.

Feature matrix initialization. We consider the pixels as nodes. The nodes feature matrix is initialized using feature

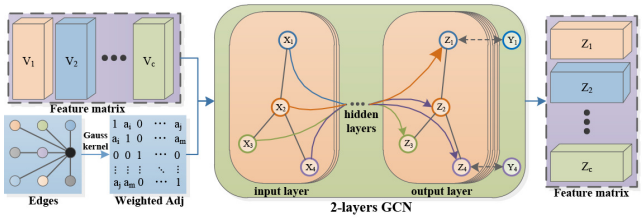

Fig. 2. Illustration of feature aggregate module(FAM) consisting of two-layers graph convolutional network.

maps from the FIMs, and the specific values are the features extracted on the local receptive field. The receptive field is the maximum range of node information received by the current node. As an example, suppose that each node is connected to its nearest L nodes, and the connections, i.e., node features, can be passed through the edges of the graph neural network. When L equals 4, for a 32×32 image, the receptive field size of the two-layers GCN is 5×32×32, which is five times larger than that of the CNN. Therefore, compared with the convolutional layers of the traditional CNN, the two-layers GCN expands the receptive field and can aggregate the features between neighboring nodes more effectively.

Adjacency matrix initialization. In the graph model, the edges are represented by the adjacency matrix. Considering the influence of distance on the correlation between nodes, the model utilizes the Gaussian kernel function to weight the adjacency matrix, i.e., the edges between the current node and neighboring nodes at different distances have various weights. The adjacency matrix after weighting by Gaussian kernel function can reflect the connection weight relationship between nodes more effectively, which is more conducive to feature transfer and feature aggregation.

The initialized feature matrix and adjacency matrix are fed into the two-layers GCN for graph convolution calcula-

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on May 11,2022 at 02:17:28 UTC from IEEE Xplore. Restrictions apply.

Table 1. Quantitative comparisons with different methods on 5 datasets with MAE (smaller is better), max/mean F-measure score (larger is better). The best results are shown in **red**.

Methods	DUTS-TE			ECSSD			HKU-IS			DUT-OMRON			PASCAL-S		
	MAE	max F	mean F	MAE	max F	mean F	MAE	max F	mean F	MAE	max F	mean F	MAE	max F	mean F
Amulet$_{17}$	0.062	0.832	0.738	0.057	0.922	0.881	0.047	0.909	0.863	0.072	0.791	0.699	0.095	0.839	0.870
NLDF$_{17}$	0.055	0.830	0.759	0.051	0.915	0.886	0.041	0.908	0.871	0.071	0.759	0.694	0.083	0.840	0.792
DSS$_{17}$	0.050	0.858	0.757	0.051	0.928	0.889	0.043	0.915	0.867	0.065	0.781	0.692	0.081	0.859	0.796
BMPM$_{18}$	0.049	0.850	0.768	0.044	0.928	0.894	0.039	0.920	0.875	0.063	0.775	0.693	0.074	0.862	0.770
PiCANet$_{18}$	0.054	0.851	0.749	0.046	0.931	0.885	0.042	0.922	0.870	0.068	0.794	0.710	0.077	0.871	0.804
PAGR$_{18}$	0.055	0.854	0.784	0.061	0.927	0.894	0.047	0.919	0.887	0.071	0.771	0.711	0.093	0.858	0.808
RAS$_{18}$	0.059	0.831	0.751	0.056	0.921	0.889	0.045	0.913	0.871	0.062	0.787	0.713	0.104	0.838	0.787
PAGE$_{19}$	0.052	0.838	0.777	0.042	0.931	0.906	0.036	0.920	0.884	0.062	0.792	0.736	0.078	0.859	0.817
HRS$_{19}$	0.051	0.843	0.793	0.054	0.920	0.902	0.042	0.913	0.892	0.066	0.762	0.708	0.090	0.852	0.809
AFNet$_{19}$	0.046	0.863	0.793	0.042	0.935	0.908	0.036	0.925	0.889	0.057	0.797	0.739	0.071	0.871	0.828
MLMSNet$_{19}$	0.049	0.852	0.745	0.045	0.928	0.868	0.039	0.920	0.871	0.064	0.774	0.692	0.075	0.864	0.771
CPD$_{19}$	0.043	0.864	0.813	0.040	0.936	0.914	0.033	0.924	0.896	0.057	0.794	0.745	0.074	0.873	0.832
PoolNet$_{19}$	0.042	0.876	0.799	0.044	0.937	0.910	0.033	0.931	0.894	0.056	0.806	0.739	0.074	0.876	0.817
EGNet$_{19}$	0.043	0.877	0.800	0.041	0.941	0.913	0.034	0.929	0.893	0.056	**0.809**	0.744	0.076	0.863	0.821
CLASS$_{20}$	0.039	-	0.833	0.038	-	0.917	0.031	-	0.909	0.057	-	0.749	0.062	-	0.838
Gatenet$_{20}$	0.045	0.870	0.783	0.041	0.941	0.896	0.036	0.929	0.889	0.061	0.794	0.723	0.071	0.880	0.808
CAGNet-V$_{20}$	0.044	0.851	0.823	0.042	0.930	0.914	0.033	0.905	0.906	0.057	0.782	0.744	0.079	0.859	0.828
DFNet-V$_{20}$	0.045	0.852	0.824	0.040	0.933	0.919	0.033	0.921	0.906	0.057	0.784	0.751	0.075	0.837	0.803
PFPN$_{20}$	0.042	0.868	0.836	0.040	0.938	0.915	0.035	0.928	0.902	0.063	0.777	0.753	0.071	**0.891**	**0.866**
MINet$_{20}$	0.039	0.877	0.822	0.038	0.942	0.921	0.031	0.931	0.904	0.056	0.793	0.743	0.064	0.868	0.828
MGCINet(Ours)	**0.038**	**0.883**	**0.838**	**0.035**	**0.945**	**0.925**	**0.029**	**0.934**	**0.911**	**0.055**	0.801	**0.754**	**0.061**	0.879	0.846

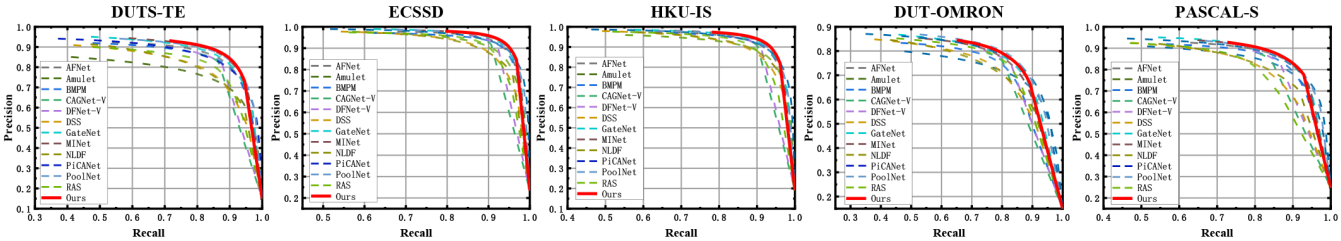

Fig. 3. Precision-Recall curves on five saliency datasets. The area below the curve indicates the model performance (lager is better). The curves show that ours (red lines) exhibit the most excellent performance.

tion. The updated feature matrix is converted into image feature maps and fed into the residual refinement module for optimization.

2.3. Residual Refinement Module

The residual refinement module (RRM) adopts a residual encoder-decoder structure, as shown in Fig.1(c). The coarse saliency maps obtained by FAMs suffer from blurred boundary segmentation. Therefore, we design the RRMs to optimize the boundary defects. The RRM is designed as a U-net residual block, which improves the coarse saliency maps by learning the residuals between the saliency predictions and the ground truth. The saliency maps after RRMs processing can acquire clearer boundaries based on refined region.

Next, every RRM is followed by a fully connected layer, which consists of a convolutional layer, a batch normalization layer and a ReLU layer. The saliency maps obtained by RRM are integrated by the fully connected layer and fed into the shallower layer.

2.4. Multi-scale Weighted Structural Loss

Our training loss is defined as the summation over all outputs: $L = \sum_{k=1}^{K} \alpha_k l^k$. Where l^k is the loss of the k-th side output, K denotes the total number of the outputs and α_k is the weight of each loss. Our multi-scale weighted structural loss is defined as:

$$l^k = l^k_{wBCE} + l^k_{wIoU} + l^k_{wSSIM} \quad (1)$$

where, l^k_{wBCE},l^k_{wIoU},and l^k_{wSSIM} denote the weighted binary cross entropy loss, the weighted intersection over union loss and the weighted structural similarity loss, respectively.

The BCE loss and the IoU loss treat pixels separately and equally during the calculation, thus ignoring the overall image structure. The SSIM loss doesn't take into account the dramatic changes in the mean and variance of the images. Therefore, the model reconstruct three weighted loss functions as multi-scale weighted structural loss to supervise the training process. It can assign distinct weights to pixels while focusing on the overall image structure at various scales.

3. EXPERIMENT

3.1. Implementation details

In this section, we will introduce some experimental details, including the datasets employed by the model, the evaluation metrics, and the training details.

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on May 11,2022 at 02:17:28 UTC from IEEE Xplore. Restrictions apply.

Table 2. Ablation analysis on PASCAL-S and DUTS-TE datasets. The better results are shown in **bold**. The best results are shown in **red**.

Settings	PASCAL-S			DUTS-TE		
	MAE	max F	mean F	MAE	max F	mean F
Baseline	0.064	0.868	0.828	0.039	0.877	0.822
+FAMs	0.062	0.873	0.839	0.038	0.880	0.828
+RRMs	0.063	0.870	0.832	0.039	0.879	0.825
+FAMs+RRMs	**0.062**	**0.875**	**0.841**	**0.038**	**0.881**	**0.830**
$+l_{wBCE}$	0.063	0.872	0.833	0.039	0.879	0.825
$+l_{wIoU}$	0.063	0.871	0.831	0.039	0.878	0.823
$+l_{wSSIM}$	0.063	0.869	0.830	0.039	0.877	0.824
$+l_{wBCE}+l_{wIoU}$	0.062	0.873	0.835	0.038	0.881	0.828
$+l_{wBCE}+l_{wSSIM}$	0.062	0.872	0.834	0.039	0.881	0.826
$+l_{wBCE}+l_{wIoU}+l_{wSSIM}$	**0.061**	**0.874**	**0.839**	**0.038**	**0.882**	**0.830**
+FAMs+RRMs+$l_{wBCE}+l_{wIoU}+l_{wSSIM}$	**0.061**	**0.879**	**0.846**	**0.038**	**0.883**	**0.838**

Datasets. We evaluate our model on five benchmark datasets. ECSSD [5] contains 1000 images containing semantic information but with complex structure. HKU-IS [8] contains 4447 images, most of which contain multiple independent salient objects or low-contrast objects close to image boundaries. DUT-OMRON [9] is composed of 5168 images with complex foreground structure. PASCAL-S [10] contains of 850 images with cluttered backgrounds and complex foregrounds objects. DUTS [11] is the largest dataset for salient object detection available, containing 10533 training images (DUTS-TR) and 5019 test images (DUTS-TE).

Evaluation metrics. We utilize four evaluation metrics to evaluate our model: mean absolute error (MAE), precision-recall (PR) curve, maximum F-measure (max F), and mean F-measure (mean F). Precision is the percentage of correctly marked saliency pixels in the predicted saliency maps. Recall is the proportion of correctly labeled saliency pixels in the ground truth. The threshold method is chosen to plot the paired sequences of precision and recall into PR curves.

Training. In the training stage, we employ the DUTS-TR as the training dataset. And data enhancement techniques such as random horizontal flipping and random rotation are utilized to avoid over-fitting problems. To ensure the convergence of the model, the network is trained on NVIDIA TITAN Xp GPU with 60 epochs at batch size 4. The backbone network in the base model is initialized using the VGG16 [12] model pre-trained on the ImageNet dataset. The remaining parameters are initialized by default using PyTorch 1.5. The model utilizes stochastic gradient descent (SGD) as the optimizer. The weight decay is set to 5e-4, the initial learning rate is 1e-3, and the momentum is set to 0.9. The input image size is 320×320.

3.2. Comparison

To verify the validity of the model, we conduct comparison experiments with the state-of-the-art SOD methods [7, 6, 13, 14, 15, 16, 17, 18, 19, 20, 21]. To be fair, the comparisons are made with the salient detection results provided by the authors or by running their publicly available models.

Quantitative comparison. Table 1 shows the detailed experimental results for the three metrics on the five benchmark datasets. As shown, our method significantly outperforms the existing state-of-the-arts. Although some metrics of our model are barely satisfactory on the PASCAL-S and DUT-OMRON datasets, our MAE values on these two datasets are outstanding. In addition, Fig.3 represents the PR curves on the five datasets, and it can be seen that our model results achieve excellent performance on all of them.

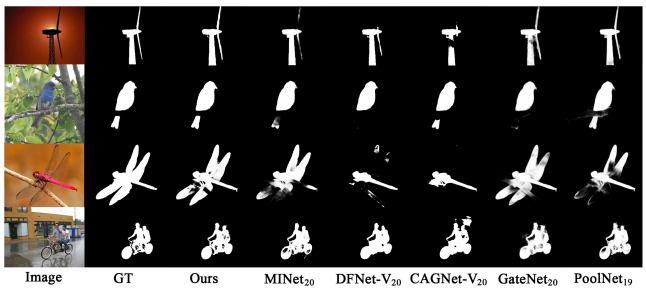

Fig. 4. Visualization comparison of different methods. Obviously, our method generates superior saliency maps.

Visualization comparison. Some representative examples of experimental results comparing with other methods are shown in Fig.4. The examples show the contrast of the saliency maps under different scenarios. From the comparison, we can see that our proposed method can acquire more complete and detailed saliency maps with clearer boundaries.

3.3. Ablation Study

To illustrate the effectiveness of each module, we perform ablation experiments on the basis of baseline, as shown in Table 2. The ablation study is mainly compared on the datasets DUTS-TE and PASCAL-S. Baseline represents the FIMs inspired by MINet [7].

At first, we individually test the validity of FAMs and RRMs based on baseline. It can be seen that both modules show superior performance compared to baseline. Next, for the combination of the two modules, the performance is further improved. Moreover, we likewise evaluate the multi-scale weighted structural loss on baseline. As shown, the combination of the three weighted loss functions exhibits the optimal results. Finally, we test the overall efficiency of FAMs and RRMs, as well as multi-scale weighted structural loss, and obviously, our proposed method shows the most excellent performance, validating the validity of the model.

4. CONCLUSION

In this paper, we propose a multi-scale graph convolutional interaction network for salient object detection. Our method can effectively aggregate information from scale-varying data by neighboring scales interaction and graph nodes interaction, while improving the boundary quality of the salient objects. Experiments show that our method significantly outperforms the state-of-the-arts on the widely used benchmark datasets.

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on May 11,2022 at 02:17:28 UTC from IEEE Xplore. Restrictions apply.

5. REFERENCES

[1] Hyemin Lee and Daijin Kim, "Salient region-based online object tracking," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 1170–1177.

[2] Yi Lu, Yaran Chen, Dongbin Zhao, and Jianxin Chen, "Graph-fcn for image semantic segmentation," in *International Symposium on Neural Networks*. Springer, 2019, pp. 97–105.

[3] Paul L Rosin and Yu-Kun Lai, "Artistic minimal rendering with lines and blocks," *Graphical Models*, vol. 75, no. 4, pp. 208–229, 2013.

[4] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu, "Global contrast based salient region detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 3, pp. 569–582, 2014.

[5] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia, "Hierarchical saliency detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 1155–1162.

[6] Bo Wang, Quan Chen, Min Zhou, Zhiqiang Zhang, Xiaogang Jin, and Kun Gai, "Progressive feature polishing network for salient object detection.," in *AAAI*, 2020, pp. 12128–12135.

[7] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu, "Multi-scale interactive network for salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9413–9422.

[8] Guanbin Li and Yizhou Yu, "Visual saliency based on multiscale deep features," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5455–5463.

[9] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3166–3173.

[10] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille, "The secrets of salient object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 280–287.

[11] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan, "Learning to detect salient objects with image-level supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 136–145.

[12] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[13] Mehrdad Noori, Sina Mohammadi, Sina Ghofrani Majelan, Ali Bahri, and Mohammad Havaei, "Dfnet: Discriminative feature extraction and integration network for salient object detection," *Engineering Applications of Artificial Intelligence*, vol. 89, pp. 103419, 2020.

[14] Sina Mohammadi, Mehrdad Noori, Ali Bahri, Sina Ghofrani Majelan, and Mohammad Havaei, "Cagnet: Content-aware guidance for salient object detection," *Pattern Recognition*, p. 107303, 2020.

[15] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang, "Suppress and balance: A simple gated network for salient object detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 35–51.

[16] Lv Tang and Bo Li, "Class: Cross-level attention and supervision for salient objects detection," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[17] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng, "Egnet: Edge guidance network for salient object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8779–8788.

[18] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang, "A simple pooling-based design for real-time salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3917–3926.

[19] Zhe Wu, Li Su, and Qingming Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3907–3916.

[20] Runmin Wu, Mengyang Feng, Wenlong Guan, Dong Wang, Huchuan Lu, and Errui Ding, "A mutual learning method for salient object detection with intertwined multi-supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8150–8159.

[21] Mengyang Feng, Huchuan Lu, and Errui Ding, "Attentive feedback network for boundary-aware salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1623–1632.

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on May 11,2022 at 02:17:28 UTC from IEEE Xplore. Restrictions apply.