Rethinking the U-shape Structure for Salient Object Detection

Jiang-Jiang Liu*, Zhi-Ang Liu*, Pai Peng, and Ming-Ming Cheng, *Senior Member, IEEE*

***Abstract*—The U-shape structure has shown its advantage in salient object detection for efficiently combining multi-scale features. However, most existing U-shape-based methods focused on improving the bottom-up and top-down pathways while ignoring the connections between them. This paper shows that we can achieve the cross-scale information interaction by centralizing these connections, hence obtaining semantically stronger and positionally more precise features. To inspire the newly proposed strategy's potential, we further design a relative global calibration module that can simultaneously process multi-scale inputs without spatial interpolation. Our approach can aggregate features more effectively while introducing only a few additional parameters. Our approach can cooperate with various existing U-shape-based salient object detection methods by substituting the connections between the bottom-up and top-down pathways. Experimental results demonstrate that our proposed approach performs favorably against the previous state-of-the-arts on five widely used benchmarks with less computational complexity. The source code will be publicly available.**

***Index Terms*—Salient object detection, U-shape structure, information interaction, deep learning**

I. Introduction

As a fundamental component of low-level computer vision and benefiting from its category-agnostic character, salient object detection has been widely applied in various downstream vision tasks, such as weakly supervised semantic segmentation [1], [2], visual tracking [3], content-aware image editing [4], and robot navigation [5]. Traditional salient object detection methods depend heavily on hand-crafted feature detectors. These detectors cannot utilize the rich high-level semantic information hidden in the image and dataset, making them fail in complex scenes. Convolutional neural networks (CNNs) based methods have been developing rapidly in recent years for their capability of extracting both high-level semantics and low-level textures of multiple scales.

Designing network architectures that can extract more expressive multi-scale features has always been a hot research direction in salient object detection, which usually brings better performance. One representative type of these architectures is the U-shape structure [6], [7]. As illustrated in the left part of Fig. 1, a typical U-shape structure consists of a bottom-up pathway, a top-down pathway, and several connections between them. Various methods have been proposed to advance the U-shape structure. The majority of these methods focus on improving the multi-scale feature extracting capability of the bottom-up pathway and/or the multi-scale feature aggregating ability of the top-down pathway. On the contrary, little attention has been paid to the connections between the two pathways. The common practice is to connect the corresponding stages between the two pathways directly or simply using a convolutional layer to map the feature channels. Thus, a natural question is whether it is possible to design an architecture that can take advantage of these previously neglected connections for multi-scale feature integration.

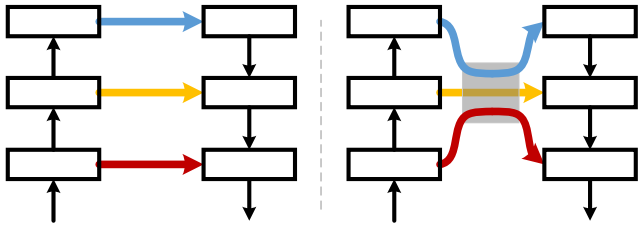


Fig. 1. Conceptual diagram of our proposed centralized information interaction strategy. **Left**: Typical U-shape structure connects the corresponding stages in the bottom-up and top-down pathways directly; **Right**: Our proposed centralized strategy. The rounded gray rectangle refers to a shared module that can parallelly process all stages (scales) of features in a stage-wise manner.

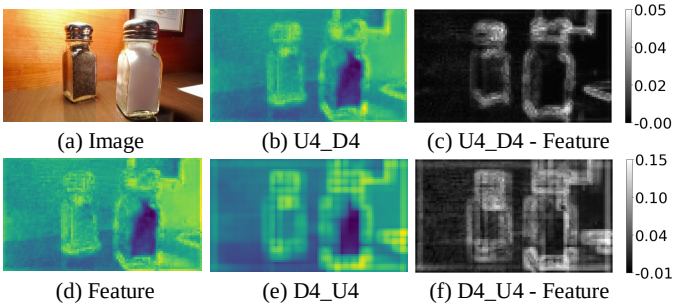
This paper focuses differently on augmenting the extracted features' representation capability by redesigning the connections between the bottom-up and top-down pathways rather than the pathways themselves. A straightforward way to achieve this purpose is to fuse (*e.g.,* concatenate or summate) the extracted multi-scale features that belong to different stages right away after the bottom-up pathway [8], [9]. However, an inevitable step in multi-scale features fusion is the spatial interpolation process, which may cause negative effects. To have an intuitive perception, we compare the intermediate feature maps before and after spatial interpolation in Fig. 2. As can be noticed, the intermediate feature map being down-sampled first and then up-sampled differs greatly from its original values and vice versa. This phenomenon worsens when the down-sampling rate increases, as more spatial information will be lost in this process.

To this end, we propose to achieve the purpose of multi-scale information interaction and obtain more expressive features by encoding the cross-scale information into the learnable filters instead of the features, where no spatial interpolation is required. As shown in the right part of Fig. 1, the multi-scale features extracted from the bottom-up pathway are parallelly processed by a centralized module whose parameters are shared across scales before being used

∗ Indicates equal contributions.

J.J. Liu, Z.A. Liu, and M.M. Cheng are with the College of Computer Science, Nankai University. M.M. Cheng is the corresponding author (cmm@nankai.edu.cn).

P. Peng is with Tencent.



(a) Image (b) U4_D4 (c) U4_D4 - Feature

(d) Feature (e) D4_U4 (f) D4_U4 - Feature

Fig. 2. An example case showing the impact of spatial interpolation: the source image (a) and the corresponding output feature map (d) of ResNet-18's `res1` stage. (b) is obtained by directly $4\times$ bilinear up-sampling and then $4\times$ bilinear down-sampling on (d), while (e) is down-sampled first and then up-sampled. (c) and (f) are the difference maps of (b,d) and (e,d), respectively. It can be seen that even the simplest spatial interpolation operations can cause obvious differences.

to build the top-down pathway. We call the proposed filter-level information interaction strategy centralized information interaction (CII). By encouraging the shared learnable filters in the centralized module to automatically adapt to the multi-scale inputs, CII implicitly ensures the cross-scale information interaction among them. CII faithfully preserves the exact spatial locations within each input scale and fundamentally avoids the negative effects caused by spatial interpolation on features. CII does not refer to any specific modules but a strategy that can cooperate with various multi-scale modules. However, existing multi-scale modules (*e.g.,* PPM [10], ASPP [11], SE [12], *etc.*) were designed to deal with single-scale input. The new normal coming along with the introduction of CII is that the shared module's inputs in CII are naturally of multiple scales. To cooperate with CII and inspire its potential, we further propose a relative global calibration module (RGC).

Unlike the previous multi-scale modules that tend to model as much multi-scale information as possible, RGC is designed to only model information at the necessary scales. RGC consists of two parallel branches, one to retain the original spatial context and the other to obtain the global embeddings w.r.t. each input scale. The embeddings are used as references to guide the feature transformation process in the original feature space for their large fields of view. Although for each single-scale input, only the original and global scales are captured, when considering the multi-scale inputs of CII, the learnable filters in RGC encode rich multi-scale information. RGC takes advantage of CII and avoids the scale diversity explosion problem [10], [11], [13] that occurs with previous multi-scale modules. We also show that with a slight modification on the input flow of the RGC module, the overall performance can be further promoted nearly for free.

To evaluate the proposed approach's performance, we report results on five popular salient object detection benchmarks. We conduct extensive ablation studies and show numerous visual examples to understand better the impacts of the proposed approach's different components. Our approach can be trained end-to-end on a single RTX-2080Ti GPU within 1.5 hours on a training set of 10,553 images. During testing, our approach can run at a speed of more than 50 FPS when processing a 300×400 image. To sum up, our main contributions can be summarized as follows:

- We rethink the U-shape structure for salient object detection by focusing on the previously neglected connections between the bottom-up and top-down pathways and propose a centralized information interaction strategy (CII). CII encodes the cross-scale information into the learnable filters instead of the features and fundamentally avoids the drawbacks caused by feature map spatial interpolation.
- We design a relative global calibration module (RGC) that utilizes the multi-scale inputting nature of CII and only models information at the necessary scales for each input scale. RGC shows the advantage and potential of CII and points a promising direction in developing multi-scale modules.
- We conduct comprehensive ablation experiments to explain the design principles and investigate the effectiveness of the proposed strategy and module.
- The proposed approach achieves superior performance over the previous state-of-the-art approaches on five challenging benchmarks with fewer parameters and FLOPs.

II. Related Work

This section briefly reviews the recent representative work on salient object detection, U-shape structures, and multi-scale and attention modules.

A. Salient Object Detection

Early salient object detection methods were usually based on intrinsic cues and hand-crafted features [14]–[17]. More details can be found in recent surveys [18]–[22]. Among the deep-learning-based methods, many adopted the idea of recurrent refinement [23], [24] to refine the predictions iteratively. Some methods treated this problem stage-wisely [25]–[27] by decoupling it into multiple stages. To get predictions with more precise boundaries, [28]–[30] designed new loss functions while [29], [31]–[33] introduced extra supervisions. [31], [34]–[38] proposed various types of attention mechanisms and achieved substantial improvements. Among the above methods, a majority of them were based on the classic U-shape structure. [38]–[42] attached additional multi-scale modules after the bottom-up pathway to generate more powerful features, while [43]–[47] combined the extracted multi-scale features in different ways within the top-down pathway to generate richer features. Regarding the expensive human annotation process, [48] proposed a novel supervision synthesis scheme combining both external and internal knowledge sources to generate supervisory signals for deep model training and obtained comparable performance.

B. U-shape Structures

How to combine the multi-scale features extracted from the backbone network has always been an attractive research direction. As pioneers, U-Net [6] and FPNs [7] were the first to incorporate an additional top-down pathway against the bottom-up pathway to sequentially combine the extracted

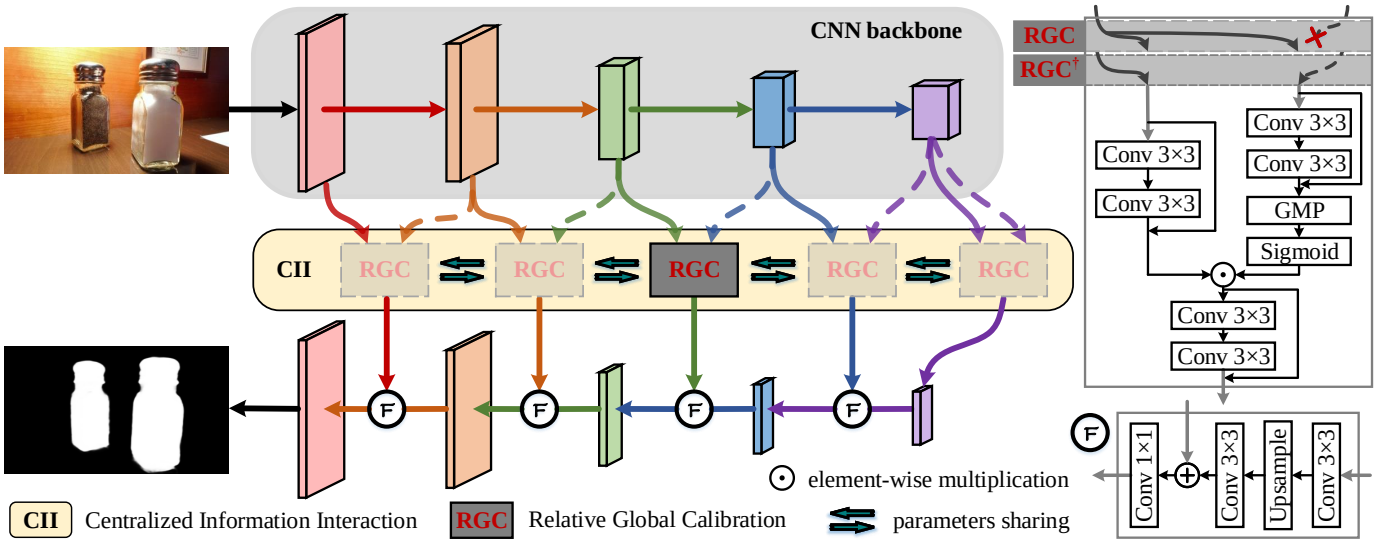


Fig. 3. The overall pipeline of our proposed approach. Note that the dashed arrows only active in **RGC**†, as shown in the top-right corner.

multi-scale features from high- to low-levels. As followers, PANet [49] adopted another bottom-up pathway on top of FPN, while ASFF [50] proposed to fuse more stages of features in the top-down pathway of FPN. EfficientDet [51] proposed a BiFPN layer and repeated it multiple times. RFP [52] proposed to repeatedly pass the features through the bottom-up backbone to enrich the representation power of FPN. Recently, NAS-FPN [53] and Auto-FPN [54] applied the neural architecture search [55] to discover the optimal FPN structure in a data-driven manner automatically.

Unlike the above methods that focused on enhancing either the feature extracting capabilities of the bottom-up pathway or the top-down pathway's feature aggregating abilities, we propose to augment the information interaction among the connections between the bottom-up and top-down pathways rather than the pathways themselves.

C. Multi-scale and Attention Modules

Deeplabv2 [11] proposed an atrous spatial pyramid pooling (ASPP) module to capture contextual information using different dilation convolutions. DenseASPP [56] improved ASPP with dense connections. PSPNet [10] utilized a pyramid pooling module (PPM) to aggregate multiple scales' contextual information with pooling operations. Recently, Auto-Deeplab [57] proved that optimal multi-scale modules could be automatically obtained with neural architecture search. CBAM [58] proposed a sequence of channel attention and spatial attention to augment the input features. OCNet [59] adopted a self-attention mechanism to augment ASPP with stronger context extraction capability. DANet [60] utilized parallelism of position attention and channel attention to model long-range dependencies. CCNet [61] developed a criss-cross attention module to capture the global context in both horizontal and vertical directions.

Unlike the modules mentioned above that use single-scale input, our proposed RGC module simultaneously takes multi-scale inputs. Our proposed CII aims to augment the information interaction among the connections between the bottom-up and top-down pathways of the U-shape structure rather than the pathways themselves.

III. Method

In this section, we first introduce the overall pipeline of the proposed approach. We then describe the two main components of the proposed approach, including an information interaction strategy and a feature calibration module.

A. Overall Pipeline

The proposed approach is based on the widely used U-shape structure, which consists of a bottom-up pathway for multi-level feature extraction and a top-down pathway to combine them. As illustrated in Fig. 3, the multi-scale features extracted from the bottom-up pathway are parallelly forwarded through the information interactors (solid gray rectangles) stage-wisely. We share these information interactors' parameters to achieve efficient cross-scale information interaction by learning more powerful filters. Then the interacted features are progressively used to build the top-down pathway from high- to low-levels. We call the above information interaction strategy that encodes the multi-scale information into the shared filters as centralized information interaction (CII). Considering that the input features of CII are now of multiple scales, we further introduce a relative global calibration (RGC) module to cooperate with it. RGC achieves a balance between essential global semantics and local textures by adaptively exploiting the relative global information concerning each different input scale. In the following subsections, we describe the strategy and module mentioned above in detail.

B. Centralized Information Interaction

One typical design in the classic U-shape structure is the short connections between the same spatial scales' stages from the bottom-up pathway to the top-down pathway. This

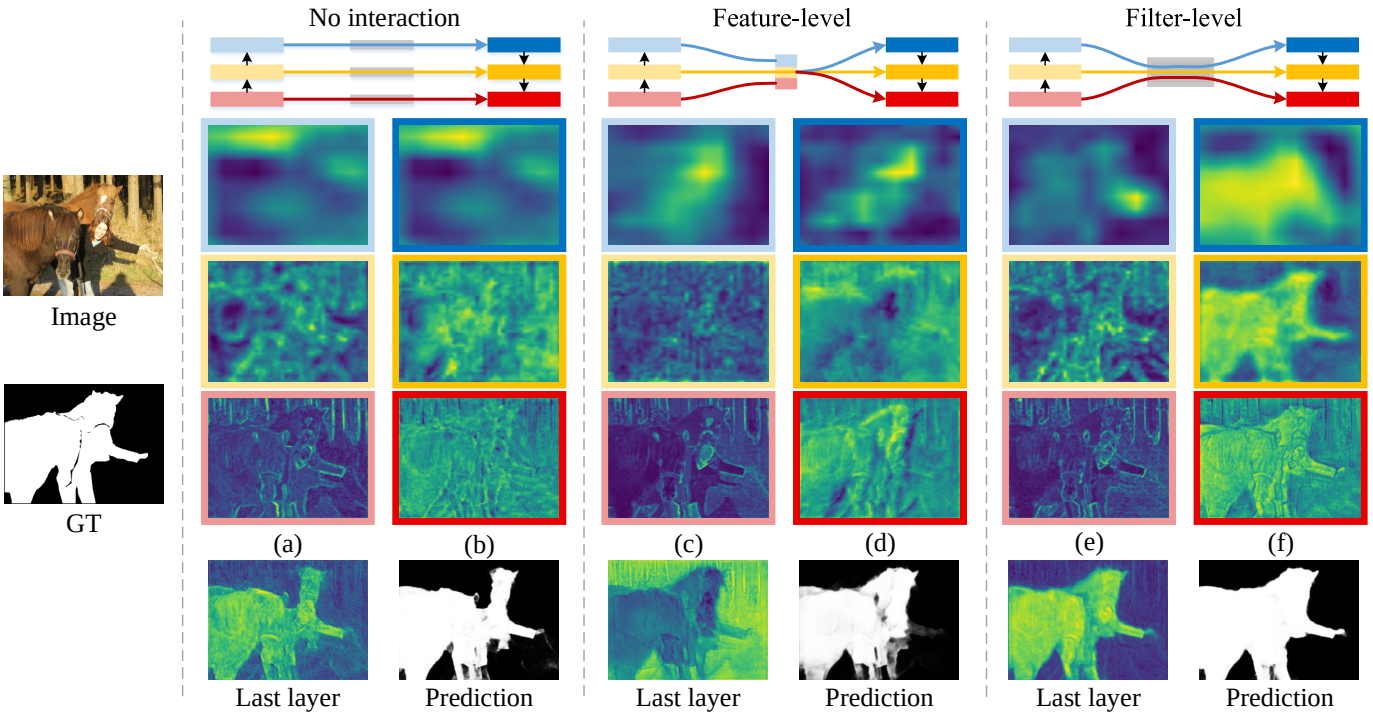


Fig. 4. Visualizing feature maps obtained with different information interaction strategies. From left to right: w/o, feature-level, and filter-level interaction. Cols (a,c,e) are features of different spatial levels before the interaction process, while Cols (b,d,f) are features after the interaction process, respectively. Subplots in the last row are features from the last convolutional layers and the corresponding predictions. As can be seen, the proposed filter-level interaction strategy (CII) can better interact the complementary information encoded in multi-scales features (Cols (f) *v.s.* (e)).

design provides a simple and efficient way to combine the extracted multi-level feature maps. We take the ResNet-18 [62] version of the classic U-shape structure as an example. The feature maps outputted by `conv1`, `res1`, `res2`, `res3`, `res4` which are denoted by $\mathbb{B} = \{B_i\}$ ($1 \leq i \leq M$ and $M = 5$) are usually used to build the output feature pyramid of the bottom-up pathway. The most high-level feature map is progressively up-sampled in the top-down pathway and then aggregated with the corresponding down-sampling rate's feature map. As can be noticed, feature map B_i can not get information from feature maps belonging to higher stages B_j ($1 \leq i < j \leq M$) until it is aggregated with B_{i+1} in the top-down pathway. Suppose we treat the connections between the bottom-up and top-down pathways as an independent part. In that case, the information flows of different scales inside this part are independent and unknown.

It is insightful that lower-level features contain more local textures and patterns while the higher-level ones indicate entire target objects' locations. It manifests the necessity of augmenting the interaction among them to complement all features with more accurate localization and more precise segmentation capabilities. To this end, we propose to enhance the information interaction among the multi-scale feature maps when they are transmitted from the bottom-up to the top-down pathways.

The classic U-shape structure directly delivers the extracted multi-scale feature maps to the top-down pathway. As shown in the rounded yellow rectangle of Fig. 3, CII differently utilizes a series of identical information interactors (solid gray rectangles) to interact with the information encoded in them. These information interactors are placed at the center of the classic U-shape structure, and the parameters of them are shared. In this way, by being encoded into the shared learnable filters, the multi-scale information can interact. Note that CII is not a specific module but a strategy. The design of the information interactors in CII is flexible and can be replaced with various successful modules [10], [11], [56], [58] to have inputs and outputs of identical shapes.

When basing on the ResNet-18 backbone, the channel numbers corresponding to $\mathbb{B}$ are set to $\{64, 64, 128, 256, 512\}$, respectively. We apply a 1×1 convolution layer ($f_i^{1\times1}$) after each $B_i \in \mathbb{B}$ to map the input channels to the same output channel (*i.e.,* 64). Note that in the following parts of the paper, we omit the batch normalization layer (BN [63]) and non-linear activation layer (ReLU [64]) after each convolution layer for notational convenience. The information interactors process the mapped feature maps to obtain the output feature maps of CII: $\mathbb{C} = \{C_i\}$ ($1 \leq i \leq M$ and $M = 5$), respectively. The overall process of CII can be summarized as:

$$C_i = InI_i(f_i^{1\times1}(B_i)),\ 1 \leq i \leq M, \tag{1}$$

where C_i and B_i are of the same spatial shapes, and InI_i refers to the identical information interactors whose parameters are shared for every i. $\mathbb{C}$ are then used to build the top-down pathway.

Quite different from the previous methods [8], [9] that achieve information interaction by directly fusing the multi-scale features (*e.g.,* concatenation or summation), our CII encodes the information into the shared learnable filters. The information interactors in CII can simultaneously get optimiza-

tion signals from high- and low-level features, resulting in semantically stronger and positionally more precise patterns. One advantage of CII is that the aliasing effect of up-sampling is fundamentally avoided. The input and output feature maps of each information interactor are of the same spatial sizes, indicating that no spatial interpolation operation is required. CII introduces only a few additional parameters since we do not need individual modules for each input scale.

To have a straightforward perception, we show some visual comparisons of the intermediate feature maps from multiple spatial levels before and after CII (Cols (e) and (f), respectively) in Fig. 4. If we leave the information interactors independent of each other (parameters not shared), we obtain a simple U-shape structure baseline (Cols (a) and (b)). By comparing Cols (e) with (f), we can see that after CII (f), the lower-level features (red borders) tend to highlight more structural areas than only the exact edge pixels. The higher-level features (blue borders), on the contrary, become more detailed and confident around the object boundaries, not just the central activation areas. As a contrast, without CII, the feature maps after and before the information interactors are visual of similar information levels (Cols (b) *v.s.* (a)). This visual phenomenon verifies the significant effect of our CII on complementing the information across scales.

C. Relative Global Calibration

CII introduces a new strategy for efficient cross-scale information interaction against the classic U-shape structure. As aforementioned, when building the information interactors, a simple sequence of two 3×3 convolution layers outperforms its baseline version with a considerable margin (the 3rd *v.s.* 2nd rows in Table. II).

It is well-known in segmentation-like tasks that an effective multi-scale module would always promote the overall performance. For example, the famous PPM [10] comprises four parallel pooling branches with different down-sampling rates to utilize the input feature maps' multi-scale information. It was first introduced in semantic segmentation and has been successfully adopted in many salient object detection methods [26], [46]. Based on this prior and that PPM is designed to be plug-and-play, we try to use it as the information interactors in CII (InI_i in Equ. (1)). However, it turns out that not only PPM but also other successful multi-scale modules (*e.g.,* ASPP [11], SE [12]) do not work well with CII (as shown in Table. VI, more details in Sec. V-C).

These existing multi-scale modules were designed to collect information from multiple receptive field sizes when the input is single-scale. When the inputs are of multiple scales (*i.e.,* $\mathbb{B}$), it results in the rapid growth of receptive field sizes (*i.e.,* 1×4 *v.s.* $M \times 4$). However, as pointed out in a lot of previous literature [10], [11], [13], more diversity does not necessarily mean better results, which can even be problematic as too much diversity may distract the following layers.

Considering that the inputs of CII (*i.e.,* $\mathbb{B}$) naturally have multiple receptive field sizes, it is essential to cut off the redundancy and retain only the necessary diversity. We propose a relative global calibration module that contains two parallel branches responsible for local information retainment and relative global information compaction, as illustrated in the top-right part of Fig. 3. Specifically, in both the two branches of RGC, $B_i \in \mathbb{B}$ is first processed by a sequence of two 3×3 convolution layers (denoted as $f_{L_2}^{3\times3}$ and $f_{R_2}^{3\times3}$, respectively). The learnable parameters in these layers leave moderate room for feature adjustment as the spatial scales of the feature maps in $\mathbb{B}$ differ. Then a global max pooling layer (GMP) is applied after the convolution layers in the right branch to compact the relative global information G_i concerning B_i:

$$G_i = \sigma(GMP((f_{R_2}^{3\times3} + 1)(B_i))), \ 1 \leq i \leq M, \quad (2)$$

where σ refers to the sigmoid function. After that, the compacted global information from the right branch is used to calibrate the retained local feature from the left branch. With another sequence of two 3×3 convolution layers ($f_{F_2}^{3\times3}$), we can obtain the output R_i:

$$R_i = (f_{F_2}^{3\times3} + 1)(G_i \odot (f_{L_2}^{3\times3} + 1)(B_i)), \ 1 \leq i \leq M. \quad (3)$$

Note that all the learnable parameters in Equ. (2) and Equ. (3) are shared for every i.

We will show in Sec. V-C that when cooperating with CII, though with fewer branches, RGC outperforms the previous multi-scale modules. To investigate the potential of RGC, we make a small modification to the inputs of RGC to lead in global information of larger receptive fields without interpolating the inputs spatially. By simply replacing the input of the right branch with the feature map from its succeeding stage (*i.e.,* B_i to $B_i + 1$), we get RGC†, which can further improve the performances while introducing no additional parameters and having even less computation complexity. We will provide more quantitative analysis in the experiment section.

IV. Experiment Setup

Datasets. For all the experiments, the DUTS-TR [66] dataset is used for training as commonly done. For performance evaluation, five popular datasets: ECSSD [67], PASCAL-S [68], DUT-OMRON [69], HKU-IS [70] and DUTS-TE [66] are used.

Loss Function. We utilize the summation of binary cross entropy (BCE) loss and intersection over union (IoU) loss as our loss function:

$$l = l_{bce} + l_{iou}. \quad (4)$$

BCE loss is broadly used in binary classification and segmentation tasks due to its robustness, which accumulates per-pixel loss in images:

$$l_{bce}(x, y) = -\frac{1}{n}\sum_{k=1}^{n}[y_k log(x_k) + (1 - y_k) log(1 - x_k)], \quad (5)$$

where x and y denote the predicted map and the ground truth, respectively, while k is the index of pixels and n is the number of pixels in x. Different from BCE loss that focuses on the pixel-level differences, IoU loss takes into account the similarity of the whole image, which is defined as follows:

$$l_{iou}(x, y) = 1 - \frac{\sum_{k=1}^{n}(y_k * x_k)}{\sum_{k=1}^{n}(y_k + x_k - y_k * x_k)}. \quad (6)$$

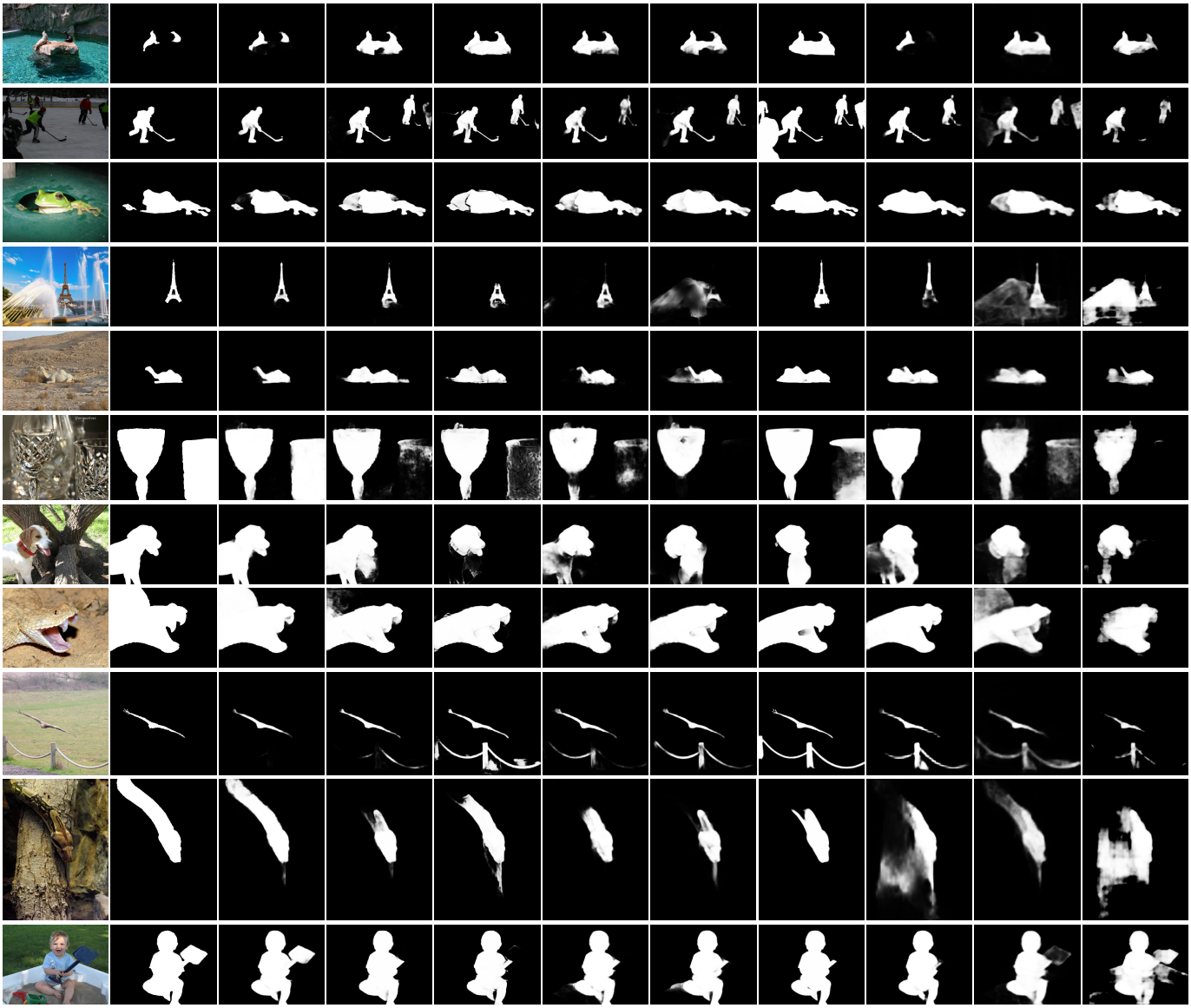


Fig. 5. Qualitative comparisons to previous state-of-the-art methods. Compared to other methods, our approach can not only locate the salient objects with cluttered backgrounds but also produce more integral saliency maps.

Evaluation Criteria. We evaluate the performance of our approach and other methods using four widely-used metrics: precision-recall (PR) curves, F-measure score (F_β), S-measure score (S_α) [71], and mean absolute error (MAE). The F-measure(F_β) score is formulated as the weighted harmonic mean of the average precision and average recall:

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}. \tag{7}$$

We set β^2 to 0.3 to weigh precision more than recall as the previous works suggested. The S-measure (S_α) score reflects both the object-aware (S_o) and the region-aware (S_r) structure similarities between the predicted map and the ground truth:

$$S_\alpha = \gamma S_o + (1-\gamma) S_r, \tag{8}$$

where γ is set as 0.5 by default. The MAE score evaluates the average pixel-level relative error between the normalized predicted map P and ground truth G:

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |P(x,y) - G(x,y)|, \tag{9}$$

where W and H denote the width and height of P, respectively.

Implementation Details. We implement our approach using the publicly available PyTorch library[1], and an RTX-2080Ti GPU is used for acceleration. The parameters of our backbone network (*i.e.,* ResNet-18 and ResNet-50 [62]) are initialized with the corresponding models pretrained on the ImageNet dataset [72], and the rest are randomly initialized. For all the experiments, our model is trained for 32 epochs with a batch size of 30. The stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and weight decay of 5e-5 is used for

[1]https://pytorch.org

TABLE I
THE COMPOSITION OF THE PROPOSED NETWORK'S PARAMETERS. AS CAN BE SEEN, THE FEATURE EXTRACTOR (RESNET-50) TAKES UP THE MAJORITY.

Total	ResNet-50	RGC	F	Others
#Params	23.51M	0.22M	0.41M	0.34M
(24.48M)	96.04%	0.90%	1.67%	1.39%

TABLE II
ABLATION ANALYSIS OF THE PROPOSED CII STRATEGY. M REFERS TO THE NUMBER OF CONNECTIONS BETWEEN THE BOTTOM-UP AND TOP-DOWN PATHWAYS. THE BEST RESULT IN EACH COLUMN IS HIGHLIGHTED IN **RED**.

Kernel Size	No.	Share	DUT-OMRON [69]			DUTS-TE [66]		
			$F_\beta \uparrow$	MAE $\downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_\alpha \uparrow$
1×1	$1 \times M$	✗	0.801	0.075	0.816	0.848	0.060	0.851
3×3	$2 \times M$	✗	0.804	0.075	0.818	0.849	0.059	0.852
3×3	2×1	✓	0.810	0.069	0.822	**0.869**	0.050	**0.870**
3×3	4×1	✓	**0.812**	**0.066**	**0.825**	0.868	**0.049**	**0.870**

optimization. The maximum learning rates are set as 0.005 for the backbone and 0.05 for the rest. We apply warm-up and cosine schedule to the learning rates in the first 8 and last 24 epochs. Random horizontal flipping and random cropping are used for data augmentation. In both training and testing, the input images are resized to 352×352. By default, our ablation experiments are based on the ResNet-18 [62] backbone unless special explanations. Note that we do not apply any pre-processing or post-processing techniques.

V. ABLATION STUDIES

In this subsection, we first conduct several straightforward experiments to show the effectiveness of the proposed CII and RGC from an overall perspective. Then we detailedly investigate the design choices and analyze the configurations of both CII and RGC with more ablation studies.

A. Composition of Parameters

We list the composition of the parameters of our network in Table. I. It can be seen that we only introduce 0.97M (3.96%) additional parameters compared to the ResNet-50 backbone (23.51M). Especially, the RGC module only occupies 0.22M (0.90%) parameters. In comparison, the other 0.75M (3.06%) parameters are essential components to construct the top-down pathway (0.41M, 1.67%) and the rest accessories of the whole network (0.34M, 1.39%). By taking advantage of the extracted multi-scale features from the backbone network and promoting the cross-scale information interaction among them, the proposed approach boosts the overall performances greatly at the cost of only a few additional parameters. The polarized composition of parameters proves the efficiency and effectiveness of the proposed approach.

B. Centralized Information Interaction

Effectiveness of CII. To demonstrate the effectiveness of CII against the classic U-shape structure, we compare different settings of the connections between the bottom-up and top-down pathways in Table. II. Except for the settings of the connections, all other configurations are identical. The 1st row in Table. II is the U-shape baseline, which connects the bottom-up and top-down pathways with a single 1×1 convolution layer, respectively. As a comparison, the 3rd row applies our proposed CII with a sequence of two 3×3 convolution layers as the information interactor (InI_i in Equ. (1)), which is shared across scales. As can be seen, the utilization of CII greatly promotes the overall performance compared to the U-shape baseline.

TABLE III
ABLATION ANALYSIS OF DIFFERENT STAGES OF FEATURES BEING USED IN CII. ✓MEANS THE CONNECTION BETWEEN THE BOTTOM-UP AND TOP-DOWN PATHWAYS, RESPECTIVELY. THE BEST RESULT IN EACH COLUMN IS HIGHLIGHTED IN **RED**.

Stage					DUT-OMRON [69]			DUTS-TE [66]		
1	2	3	4	5	$F_\beta \uparrow$	MAE $\downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_\alpha \uparrow$
✓	✓	✓			0.805	0.071	0.822	0.858	0.055	0.861
	✓	✓	✓		0.808	0.067	0.822	0.860	0.052	0.862
		✓	✓	✓	0.807	0.067	0.824	0.859	0.052	0.864
✓	✓	✓	✓		0.805	0.067	0.823	0.864	0.051	0.867
	✓	✓	✓	✓	0.808	0.067	**0.825**	0.867	0.050	0.869
✓	✓	✓	✓	✓	**0.812**	**0.066**	**0.825**	**0.868**	**0.049**	**0.870**

Influence of Additional Parameters. To eliminate the influence of introducing more parameters, we replace every 1×1 convolution layer in the model of the 1st row with a sequence of two 3×3 convolution layers (M sequences, $2 \times M$ convolution layers in total), as shown in the 2nd row of Table. II. With more convolutional layers, the overall performances are slightly improved (the 2nd *v.s.* 1st rows). However, there is still a large margin compared to CII (the 2nd *v.s.* 3rd rows). This phenomenon demonstrates that more parameters do not necessarily mean better performance. A similar conclusion can be drawn by comparing the 4th with the 3rd rows in Table. II.

Different Stages of Features. To verify how different combinations of stages of features used in CII influence performance, we conduct a series of ablation experiments in Table. III. The overall trend is that more stages of features usually bring better results. By comparing the first three rows, which only use three stages of features, we can find out that leaving out neither too much low- nor high-level features harms badly. Especially when we only take the first three stages (the 1st row), the MAE (lower is better) scores of both datasets are worse. We can also observe that lacking neither the lowest- nor highest-level features will decrease the overall performance. The best performances are reached when all five stages of features are utilized.

Filter-level *v.s.* Feature-level Interaction. We report the performance obtained with different information interaction strategies in Table. IV. We can see that the proposed filter-level information interaction strategy performs better than the feature-level one against the baseline setting (the 3rd *v.s.* 2nd *v.s.* 1st rows), especially on the more challenging DUTS-TE set. We also compare the intermediate feature maps before

TABLE IV
ABLATION ANALYSIS ON DIFFERENT INFORMATION INTERACTION STRATEGIES. THE FIRST COLUMN INDICATES THE NUMBER OF ADDITIONAL CONVOLUTIONAL LAYERS USED BETWEEN THE BOTTOM-UP AND TOP-DOWN PATHWAYS. THE BEST RESULT IN EACH COLUMN IS HIGHLIGHTED IN **RED**.

No. Convs	Strategy	DUT-OMRON [69]			DUTS-TE [66]		
		$F_\beta \uparrow$	MAE $\downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_\alpha \uparrow$
5	no interaction	0.801	0.075	0.816	0.848	0.060	0.851
5	feature-level	0.808	0.069	0.821	0.854	0.055	0.858
4	filter-level	**0.812**	**0.066**	**0.825**	**0.868**	**0.049**	**0.870**

TABLE V
ABLATION ANALYSIS OF THE NUMBER OF CONVOLUTIONAL LAYERS IN EACH BRANCH OF THE RGC MODULE. THE BEST RESULT IN EACH COLUMN IS HIGHLIGHTED IN **RED**.

# 3×3 Convs	DUT-OMRON [69]			DUTS-TE [66]		
	$F_\beta \uparrow$	MAE $\downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_\alpha \uparrow$
0×3	0.805	0.063	0.815	0.853	0.049	0.854
1×3	0.810	0.064	0.819	0.868	0.047	0.866
$\mathbf{2 \times 3}$	**0.820**	0.061	**0.826**	0.873	0.045	0.870
3×3	0.810	0.061	0.824	**0.874**	**0.044**	**0.872**
4×3	0.810	**0.060**	0.821	0.872	0.045	0.866

TABLE VI
ABLATION ANALYSIS OF THE PROPOSED RGC MODULE. THE 1ST ROW USES A SEQUENCE OF FOUR 3×3 CONVOLUTION LAYERS AS THE INFORMATION INTERACTOR (THE SAME AS THE 4TH ROW IN TABLE. II). THE BEST RESULT IN EACH COLUMN IS HIGHLIGHTED IN **RED**.

Information Interactor	DUT-OMRON [69]			DUTS-TE [66]		
	$F_\beta \uparrow$	MAE $\downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_\alpha \uparrow$
3×3 Convs	0.812	0.066	0.825	0.868	0.049	0.870
RGC	0.820	0.061	0.826	0.873	0.045	0.870
RGC†	**0.824**	**0.058**	**0.828**	**0.878**	**0.042**	**0.874**
PPM [10]	0.810	0.066	0.826	0.868	0.047	0.868
PPM†	0.803	0.070	0.821	0.866	0.047	0.869
ASPP [11]	0.814	0.06	0.820	0.872	0.047	0.864
ASPP†	0.806	0.067	0.824	0.867	0.047	0.869
SE [12]	0.809	0.066	0.823	0.869	0.046	0.870
SE†	0.813	0.063	0.827	0.872	0.045	0.869

and after applying different information interaction strategies in Fig. 4. Before information interaction (Cols (a, c, e)), the high-level features (light blue borders) mostly focus on the coarse locations of the salient objects. In contrast, the middle- (light yellow) and low-level (pink) features usually highlight the edge pixels. Without information interaction (Cols (b) *v.s.* (a)), the features of corresponding levels are roughly visually unchanged. After feature-level information interaction (Cols (d) *v.s.* (c)), the salient objects in the high-level features (blue) are slightly more distinct from the backgrounds, while the middle- (yellow) and low-level (red) features begin to highlight the integrated salient objects rather than only the edge pixels. More obvious improvement can be observed (Cols (f) *v.s.* (d,e)) when the proposed filter-level interaction strategy is applied. The activations of the salient objects in the high-level features are stronger and more integrated, while the highlights in the middle- and low-level features show preciser localization and sharper segmentation.

We argue that the above improvement largely owes that the proposed filter-level interaction strategy requires no spatial interpolation on the feature maps. The high-resolution details are furthest retained (w/o down-sampling) and avoiding involving aliasing effect (w/o up-sampling). However, for the feature-level interaction, features from all levels are interpolated and then concatenated together, where the unfaithful information brought by interpolation can be misleading and harmful. As can be seen from Col (d), the highlights in the middle- and low-level features have more blurry boundaries and less confident object centers compared with Col (f). A similar conclusion can also be drawn by comparing the features from the last convolution layers and the corresponding predictions in the last row of Fig. 4.

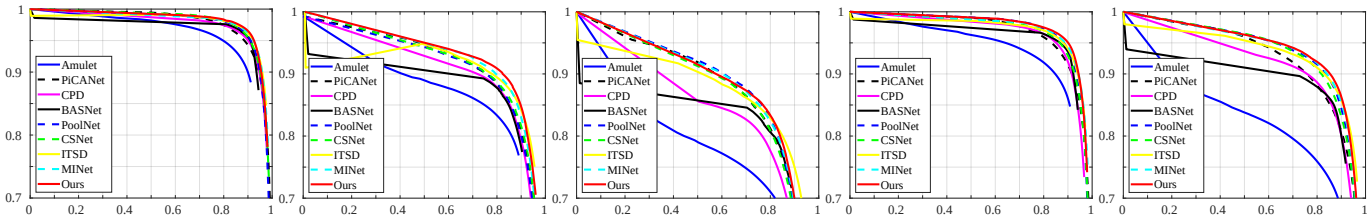
C. Relative Global Calibration

Effectiveness of RGC. To prove the effectiveness of RGC, we conduct a series of ablation experiments comparing different settings of the centralized information interactors (InI_i in Equ. (1)). Note that in the following experiments, except for the information interactor itself, all other configurations are identical to the 4th row in Table. II. As can be seen from the 2nd row compared to the 1st row in Table. VI, with the help of the RGC module, better overall performances are reached. It proves the additional branch's essentiality in RGC, which calibrates the local feature with its relative global information within each specific input scale.

The Number of Convolutional Layers. We conduct a series of experiments to determine the optimal number of convolutional layers in Table. V. As can be seen from the top three rows, the overall performances go up when more convolutional layers are introduced. However, if we try to increase the number of convolutional layers further, the overall performances do not rise but decrease slightly (the last three rows). We choose the 3rd row (each branch in the RGC module contains two 3×3 convolutional layers) as our default setting, which reaches a balance between effectiveness and efficiency. We also illustrate this setting in the top-right corner of Fig. 3.

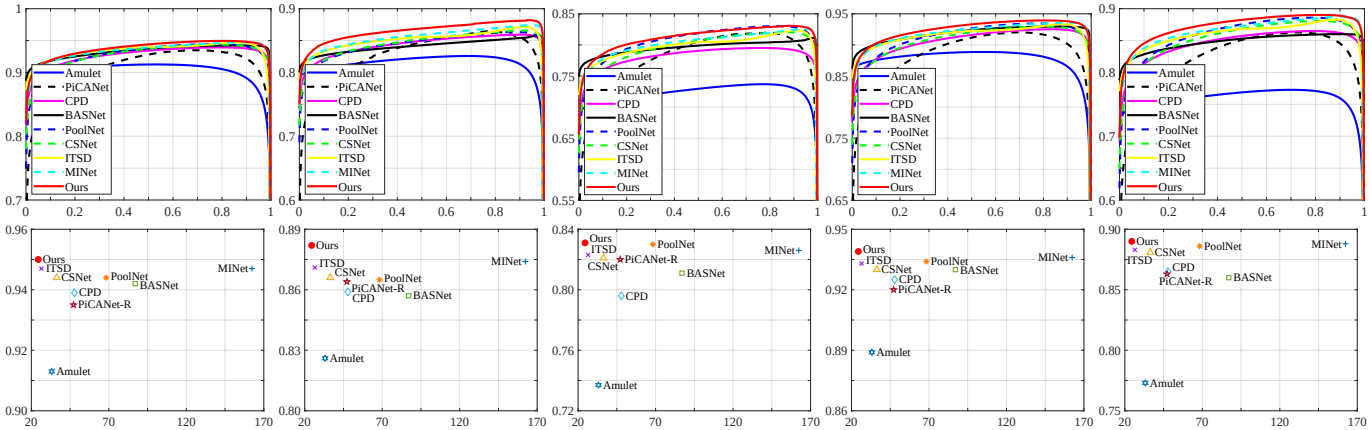
More Global Information. RGC† changes the input of its right (relative global) branch to be the succeeding stage's feature (B_i to B_{i+1}). This modification of input flow requires no additional parameters nor computational burdens. By comparing the 3rd to 2nd rows in Table. VI, we can see that the performances are further promoted. It is only an example showing the potential of designing new modules and strategies better working with the proposed CII. We hope our design principles and experiments could provide future research with more insights.

Comparison to Previous Modules. To investigate how the previous modules work with the proposed CII, we migrate some representative and successful modules (*e.g.,* PPM [10], ASPP [11], and SE [12]) into CII by directly replacing the RGC module with it. However, as shown in Table. VI, neither of these previous successful modules performs well. The PPM and SE modules even perform slightly worse than those with



(a) ECSSD [67] (b) PASCAL-S [68] (c) DUT-OMRON [69] (d) HKU-IS [70] (e) DUTS-TE [66]

Fig. 6. Precision (vertical axis) recall (horizontal axis) curves on five popular salient object detection datasets.



(a) ECSSD [67] (b) PASCAL-S [68] (c) DUT-OMRON [69] (d) HKU-IS [70] (e) DUTS-TE [66]

Fig. 7. The first row shows the F-measure (vertical axis) and threshold (horizontal axis) curves. The second row shows the comparison of F-measure scores (vertical axis) and the number of parameters (horizontal axis).

only 3×3 convolution layers (the 1st row). We also try the † versions of these three modules (*i.e.,* PPM†, ASPP†, and SE†), which are obtained by changing the inputs of the branches with spatial interpolation (for PPM and SE) or dilation convolutional (for ASPP) operations to be the features of the succeeding stage. Not surprisingly, the results are not satisfactory (the 5th, 7th, and 9th rows in Table. VI). An interesting phenomenon is that the performances of PPM† and ASPP† decrease apparently, while the ones of SE† increase. These numerical results indicate that the previous successful modules may not necessarily succeed in CII. It shows a promising direction of developing new multi-scale modules better cooperating with CII.

VI. Comparisons to the State-of-the-Arts

In this section, we compare our proposed approach with 26 previous state-of-the-art methods, including DCL [73], RFCN [74], WSS [66], MSR [34], DSS [43], NLDF [45], Amulet [44], SRM [26], C2SNet [75], PAGR [36], DGRL [23], RAS [76], PiCANet [35], AFNet [31], MLMS [77], JDFPR [78], PAGE [37], CapSal [79], ICTB [24], CPD [27], BASNet [28], PoolNet [46], CSNet [65], GateNet [42], ITSD [29], and MINet [47]. For fair comparisons, the saliency maps of other methods are generated by the original codes released by the corresponding authors or directly provided by them. We evaluate all the results with the same evaluation codes.

Computational Complexities and Numerical Scores. Quantitative results are listed in Table. VII. As can be seen, the ResNet-50 version of our proposed approach achieves the best performances on most of the datasets and metrics while requiring fewer parameters and FLOPs than the previous good-performing methods on the same backbone network. When compared to the previous most efficient method ITSD [29] (based on ResNet-50), our method achieves obvious better results on most of the datasets while requiring 67% fewer additional parameters (0.97M *v.s.* 2.96M) and 14% fewer additional FLOPs (4.74G *v.s.* 5.53G)

We also show the results of the ResNet-18 version of our approach in Table. VII. Surprisingly, our ResNet-18 version still outperforms most of the previous methods based on the more powerful ResNet-50 network with even fewer parameters and FLOPs. This phenomenon shows the effectiveness of the proposed method in taking advantage of the extracted multi-scale features by efficiently accelerating the information interaction process.

PR and F-measure Curves. Besides the numerical comparison, we also show the PR and F-measure curves on the five datasets in Fig. 6 and Fig. 7, respectively. It can be seen that the PR and F-measure curves of our approach (red ones) are comparable to other previous methods on most of the datasets and even better on some datasets. Especially on the challenging PASCAL-S dataset, our approach surpasses almost all other methods under most thresholds, which is praiseworthy considering it only introduces 3.96% additional parameters besides the backbone network. Our approach achieves better performances while requiring fewer parameters (solid red dots in the 2nd row of Fig. 7). It verifies the compactness and

TABLE VII
QUANTITATIVE COMPARISONS ON FIVE WIDELY USED DATASETS. THE BEST AND SECOND-BEST RESULTS IN EACH COLUMN ARE HIGHLIGHTED IN **RED** AND **BLUE**, RESPECTIVELY. WE ALSO SHOW THE RESULTS OF THE RESNET-18 VERSION OF OUR APPROACH. LINES IN GRAY OR BLUE MEAN METHODS OF SIMILAR COMPUTATIONAL COMPLEXITIES (FLOPS), RESPECTIVELY. THE FLOPS OF ALL APPROACHES ARE MEASURED WITH AN INPUT IMAGE SIZE OF 224×224.

Method	Params (M)	FLOPs (G)	ECSSD [67]			PASCAL-S [68]			DUT-OMRON [69]			HKU-IS [70]			DUTS-TE [66]		
			$F_\beta \uparrow$	MAE $\downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_\alpha \uparrow$
DCL_{16} [73]	66.25	-	0.898	0.078	0.873	0.805	0.115	0.800	0.733	0.094	0.762	0.893	0.063	0.871	0.786	0.081	0.803
$RFCN_{16}$ [74]	-	-	0.898	0.095	0.860	0.827	0.118	0.808	0.747	0.094	0.774	0.895	0.079	0.860	0.786	0.090	0.793
WSS_{17} [66]	-	-	0.855	0.104	0.811	0.771	0.140	0.740	0.694	0.110	0.726	0.862	0.079	0.819	0.740	0.099	0.743
MSR_{17} [34]	-	-	0.906	0.056	0.892	0.839	0.083	0.835	0.790	0.073	0.805	0.907	0.043	0.896	0.824	0.062	0.834
DSS_{17} [43]	62.23	52.20	0.908	0.062	0.884	0.821	0.101	0.804	0.760	0.074	0.789	0.900	0.050	0.881	0.813	0.065	0.826
$NLDF_{17}$ [45]	35.48	-	0.905	0.063	0.875	0.822	0.098	0.805	0.753	0.079	0.770	0.902	0.048	0.878	0.816	0.065	0.816
$Amulet_{17}$ [44]	33.16	20.70	0.913	0.060	0.881	0.826	0.092	0.816	0.737	0.083	0.784	0.889	0.052	0.866	0.773	0.075	0.800
SRM_{17} [26]	53.14	-	0.917	0.054	0.895	0.838	0.084	0.834	0.769	0.069	0.798	0.906	0.046	0.887	0.826	0.058	0.836
$C2SNet_{18}$ [75]	-	-	0.910	0.055	0.894	0.842	0.082	0.836	0.757	0.072	0.798	0.896	0.048	0.883	0.807	0.062	0.828
$PAGR_{18}$ [36]	-	-	0.927	0.061	0.889	0.847	0.089	0.822	0.771	0.071	0.775	0.919	0.047	0.889	0.854	0.055	0.839
$DGRL_{18}$ [23]	-	-	0.922	0.041	0.903	0.844	0.072	0.836	0.774	0.062	0.806	0.910	0.036	0.895	0.828	0.049	0.842
RAS_{18} [76]	20.23	21.24	0.921	0.056	0.893	0.829	0.101	0.799	0.786	0.062	0.814	0.913	0.045	0.887	0.831	0.059	0.839
$PiCANet_{18}$ [35]	47.22	54.06	0.935	0.047	0.917	0.864	0.075	0.854	0.820	0.064	0.830	0.920	0.044	0.904	0.863	0.050	0.868
$AFNet_{19}$ [31]	25.78	-	0.936	0.042	0.914	0.861	0.070	0.849	0.820	0.057	0.825	0.926	0.036	0.906	0.867	0.045	0.867
$MLMS_{19}$ [77]	74.38	58.18	0.930	0.045	0.911	0.853	0.074	0.844	0.793	0.063	0.809	0.922	0.039	0.907	0.854	0.048	0.862
$JDFPR_{19}$ [78]	87.61	42.96	0.928	0.049	0.907	0.854	0.082	0.841	0.802	0.057	0.821	-	-	-	0.833	0.058	0.836
$PAGE_{19}$ [37]	-	-	0.931	0.042	0.912	0.848	0.076	0.842	0.791	0.062	0.825	0.920	0.036	0.904	0.838	0.051	0.855
$CapSal_{19}$ [79]	-	-	-	-	-	0.862	0.073	0.837	-	-	-	0.889	0.058	0.851	0.844	0.060	0.818
$ICTB_{19}$ [24]	-	-	0.938	0.041	0.918	0.855	0.071	0.850	0.811	0.060	0.837	0.925	0.037	0.909	0.855	0.043	0.865
CPD_{19} [27]	47.85	7.23	0.939	0.037	0.918	0.859	0.071	0.848	0.796	0.056	0.825	0.925	0.034	0.907	0.865	0.043	0.869
$BASNet_{19}$ [28]	87.06	97.65	0.942	0.037	0.916	0.857	0.076	0.838	0.811	0.057	0.836	0.930	0.033	0.908	0.860	0.047	0.866
Ours(ResNet-18)	11.89	6.49	0.941	0.039	0.916	0.868	0.068	0.851	0.824	0.058	0.828	0.933	0.032	0.912	0.878	0.042	0.874
$PoolNet_{19}$ [46]	68.26	38.19	0.944	0.039	0.921	0.865	0.075	0.850	0.830	**0.055**	0.836	0.934	0.032	0.917	0.886	0.040	0.883
$CSNet_{20}$ [65]	36.37	11.75	0.944	0.038	0.921	0.866	0.073	0.851	0.821	**0.055**	0.831	0.930	0.033	0.911	0.881	0.040	0.879
$GateNet_{20}$ [42]	128.63	55.23	0.946	0.040	0.920	**0.877**	0.068	0.858	**0.831**	**0.055**	0.838	0.935	0.033	0.915	**0.889**	0.040	**0.885**
$ITSD_{20}$ [29]	26.47	9.67	**0.947**	**0.035**	**0.925**	0.871	0.066	**0.859**	0.823	0.061	**0.840**	0.933	0.031	0.916	0.883	0.041	**0.885**
$MINet_{20}$ [47]	162.38	42.73	**0.947**	**0.034**	**0.925**	0.874	**0.064**	0.856	0.826	0.056	0.833	**0.936**	**0.028**	**0.920**	0.888	**0.037**	0.884
Ours(ResNet-50)	24.48	8.88	**0.950**	**0.034**	**0.926**	**0.882**	**0.062**	**0.865**	**0.831**	**0.054**	**0.839**	**0.939**	**0.029**	**0.920**	**0.890**	**0.036**	**0.888**

robustness of our approach.

Visual Comparisons. In Fig. 5, we show some representative examples to evaluate our proposed method visually. It can be easily observed that our proposed method can not only accurately highlight the salient objects but also segment them out integrally in almost all circumstances. Different from [32], [37], [80], [81], our approach requires no extra supervision on the edge areas. It demonstrates our method's effectiveness in augmenting the information interaction among the multi-scale features. Our method can generate semantically stronger and positionally more precise features.

Failure Case Analysis. Some typical failure predictions of our approach are shown in Fig. 8. After looking through all the failure cases, we find that most of them can be roughly categorized into three circumstances: complex background, occlusion, and low contrast between foreground and background. As shown in the top two rows, these examples all have cluttered backgrounds around the salient objects. In the middle two rows, some irrelevant objects occlude part of the salient objects. Moreover, in the last two rows, some salient objects have very similar colors compared to their background. In some cases, only part of the salient objects is detected by our approach. In contrast, in other cases, the non-salient regions belong to the background are wrongly predicted as salient objects. We argue that it is hard for humans to precisely distinguish the boundaries between the foreground and background in most circumstances.

Cooperation with Other U-shape Structures. To investigate the generalization ability of the proposed CII and RGC, we apply them to two representative U-shape-based salient object detection methods: BASNet [28] and PoolNet [46]. Since the proposed CII and RGC do not rely on the specific implementation forms of the bottom-up and top-down pathways, we only replace the previously identical connections between them. The numerical results are shown in Table. VIII. As can be seen, cooperating the proposed CII strategy with either one of the two methods promotes performance. An interesting phenomenon is that the proposed CII strategy brings more improvements to BASNet than PoolNet (average promotions of 3.05% *v.s.* 0.45% in terms of F_β). We think it may be because PoolNet empirically guides the global information to each stage of the top-down pathway, already

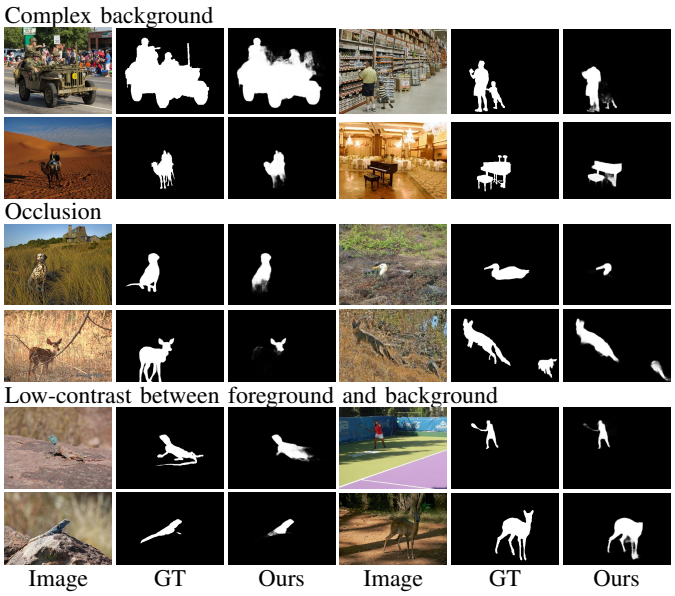


Fig. 8. Typical failure cases selected from multiple datasets.

TABLE VIII
ABLATION ANALYSIS OF THE GENERALIZATION ABILITY OF CII AND RGC TO OTHER U-SHAPE MODELS. THE BEST RESULT IN EACH COLUMN IS HIGHLIGHTED IN **RED**.

Method	CII	RGC	DUT-OMRON [69]			DUTS-TE [66]		
			$F_\beta \uparrow$	MAE $\downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_\alpha \uparrow$
BASNet [28]	✗	✗	0.811	0.057	0.836	0.860	0.047	0.866
	✓	✗	0.837	**0.053**	0.833	0.890	0.037	0.885
	✓	✓	**0.839**	0.054	**0.840**	**0.895**	**0.036**	**0.891**
PoolNet [46]	✗	✗	0.830	0.055	0.836	0.886	0.040	0.883
	✓	✗	0.831	0.054	0.838	0.894	0.035	0.891
	✓	✓	**0.839**	**0.052**	**0.842**	**0.900**	**0.034**	**0.892**

achieving moderate cross-scale information interaction. On the contrary, BASNet only uses the basic U-shape structure, so the introduction of CII leads to more improvements. When the RGC module is introduced, the overall performance of both methods can be further promoted on nearly all evaluation criteria and datasets, setting up new state-of-the-arts. The above experiments show that the proposed CII and RGC complement methods that focus on improving the bottom-up or/and top-down pathways of the U-shape structures.

VII. CONCLUSIONS AND FUTURE WORK

This paper advances the classic U-shape structure by centralizing the previous independent connections between its bottom-up and top-down pathways to encourage the information interaction among multi-scale features. To show that this centralized information interaction (CII) strategy is feasible, we propose a relative global calibration (RGC) module to cooperate with it. Combining CII and RGC into the classic U-shape architecture shows that our proposed approach can surpass all previous state-of-the-art methods on five widely-used salient object detection benchmarks. We only introduce a handful number of additional parameters and FLOPs. Our proposed strategy and module are independent of the bottom-up and top-down pathways' designs in the U-shape structures and can be flexibly applied to any U-shape based models. The proposed RGC shows a promising direction of developing new multi-scale modules better cooperating with the proposed CII strategy.

ACKNOWLEDGMENT

This research was supported by the Major Project for New Generation of AI under Grant No. 2018AAA0100400, NSFC (61922046), and S&T innovation project from Chinese Ministry of Education.

REFERENCES

[1] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, and S. Yan, "Stc: A simple to complex framework for weakly-supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016.
[2] Q. Hou, P.-T. Jiang, Y. Wei, and M.-M. Cheng, "Self-erasing network for integral object attention," in *Adv. Neural Inform. Process. Syst.*, 2018.
[3] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," in *Int. Conf. Mach. Learn.*, 2015, pp. 597–606.
[4] M.-M. Cheng, F.-L. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Repfinder: finding approximately repeated scene elements for image editing," *ACM Trans. Graphics*, vol. 29, no. 4, p. 83, 2010.
[5] C. Craye, D. Filliat, and J.-F. Goudou, "Environment exploration for object-based visual saliency learning," in *ICRA*, 2016, pp. 2303–2309.
[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241.
[7] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection." in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.
[8] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra r-cnn: Towards balanced learning for object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 821–830.
[9] Z. Li, C. Lang, J. Liew, Q. Hou, Y. Li, and J. Feng, "Cross-layer feature pyramid network for salient object detection," *arXiv preprint arXiv:2002.10864*, 2020.
[10] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.
[11] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017.
[12] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 7132–7141.
[13] J.-J. Liu, Q. Hou, M.-M. Cheng, C. Wang, and J. Feng, "Improving convolutional networks with self-calibrated convolutions," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.
[14] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 733–740.
[15] J. Wang, H. Jiang, Z. Yuan, M.-M. Cheng, X. Hu, and N. Zheng, "Salient object detection: A discriminative regional feature integration approach," *International Journal of Computer Vision*, vol. 123, no. 2, pp. 251–268, 2017.
[16] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, 2015.
[17] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *Int. Conf. Comput. Vis.*, 2013, pp. 2976–2983.
[18] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, 2015.
[19] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," *Computational Visual Media*, vol. 5, no. 2, pp. 117–150, 2019.
[20] W. Wang, J. Shen, J. Xie, M.-M. Cheng, H. Ling, and A. Borji, "Revisiting video saliency prediction in the deep learning era," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 220–237, 2021.

[21] T. Zhou, D.-P. Fan, M.-M. Cheng, J. Shen, and L. Shao, "Rgb-d salient object detection: A survey," *Computational Visual Media*, pp. 1–33, 2021.

[22] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: a survey," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 84–100, 2018.

[23] T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, and A. Borji, "Detect globally, refine locally: A novel approach to saliency detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 3127–3135.

[24] W. Wang, J. Shen, M.-M. Cheng, and L. Shao, "An iterative and cooperative top-down and bottom-up inference network for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[25] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[26] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, "A stagewise refinement model for detecting salient objects in images," in *Int. Conf. Comput. Vis.*, 2017, pp. 4019–4028.

[27] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[28] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[29] H. Zhou, X. Xie, J.-H. Lai, Z. Chen, and L. Yang, "Interactive two-stream decoder for accurate and fast saliency detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 9141–9150.

[30] J. Wei, S. Wang, and Q. Huang, "F3net: Fusion, feedback and focus for salient object detection," in *The National Conference on Artificial Intelligence*, 2020.

[31] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[32] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "Egnet:edge guidance network for salient object detection," in *Int. Conf. Comput. Vis.*, 2019.

[33] J. Wei, S. Wang, Z. Wu, C. Su, Q. Huang, and Q. Tian, "Label decoupling framework for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 13 025–13 034.

[34] G. Li, Y. Xie, L. Lin, and Y. Yu, "Instance-level salient object segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[35] N. Liu, J. Han, and M.-H. Yang, "Picanet: Learning pixel-wise contextual attention for saliency detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 3089–3098.

[36] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 714–722.

[37] W. Wang, S. Zhao, J. Shen, S. C. H. Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[38] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[39] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang, "A bi-directional message passing model for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 1741–1750.

[40] J. Su, J. Li, Y. Zhang, C. Xia, and Y. Tian, "Selectivity or invariance: Boundary-aware salient object detection," in *Int. Conf. Comput. Vis.*, 2019, pp. 3799–3808.

[41] Y. Zeng, P. Zhang, J. Zhang, Z. Lin, and H. Lu, "Towards high-resolution salient object detection," in *Int. Conf. Comput. Vis.*, 2019, pp. 7234–7243.

[42] X. Zhao, Y. Pang, L. Zhang, H. Lu, and L. Zhang, "Suppress and balance: A simple gated network for salient object detection," in *Eur. Conf. Comput. Vis.*, 2020.

[43] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 815–828, 2019.

[44] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Int. Conf. Comput. Vis.*, 2017.

[45] Z. Luo, A. K. Mishra, A. Achkar, J. A. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection." in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[46] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[47] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 9413–9422.

[48] D. Zhang, J. Han, Y. Zhang, and D. Xu, "Synthesizing supervision for learning deep saliency network without human annotation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 7, pp. 1755–1769, 2019.

[49] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 8759–8768.

[50] S. Liu, D. Huang, and Y. Wang, "Learning spatial fusion for single-shot object detection," *arXiv preprint arXiv:1911.09516*, 2019.

[51] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 10 781–10 790.

[52] S. Qiao, L.-C. Chen, and A. Yuille, "Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution," *arXiv preprint arXiv:2006.02334*, 2020.

[53] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "Nas-fpn: Learning scalable feature pyramid architecture for object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 7036–7045.

[54] H. Xu, L. Yao, W. Zhang, X. Liang, and Z. Li, "Auto-fpn: Automatic network architecture adaptation for object detection beyond classification," in *Int. Conf. Comput. Vis.*, 2019, pp. 6649–6658.

[55] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," in *Int. Conf. Learn. Represent.*, 2017.

[56] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "Denseaspp for semantic segmentation in street scenes," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 3684–3692.

[57] C. Liu, L.-C. Chen, F. Schroff, H. Adam, W. Hua, A. L. Yuille, and L. Fei-Fei, "Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 82–92.

[58] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[59] Y. Yuan and J. Wang, "Ocnet: Object context network for scene parsing," *arXiv preprint arXiv:1809.00916*, 2018.

[60] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 3146–3154.

[61] Z. H. . X. W. . Y. W. . L. H. . H. S. . W. L. . T. S. Huang, "Ccnet: Criss-cross attention for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2020.

[62] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[63] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Int. Conf. Mach. Learn.*, 2015.

[64] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Int. Conf. Mach. Learn.*, 2010.

[65] S.-H. Gao, Y.-Q. Tan, M.-M. Cheng, C. Lu, Y. Chen, and S. Yan, "Highly efficient salient object detection with 100k parameters," in *Eur. Conf. Comput. Vis.*, 2020.

[66] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 136–145.

[67] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 1155–1162.

[68] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 280–287.

[69] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 3166–3173.

[70] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 5455–5463.

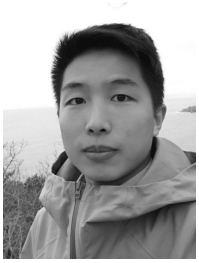
[71] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A New Way to Evaluate Foreground Maps," in *Int. Conf. Comput. Vis.*, 2017, pp. 4548–4557.

[72] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Adv. Neural Inform. Process. Syst.*, 2012.

[73] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[74] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *Eur. Conf. Comput. Vis.*, 2016.

[75] X. Li, F. Yang, H. Cheng, W. Liu, and D. Shen, "Contour knowledge transfer for salient object detection," in *Eur. Conf. Comput. Vis.*, 2018, pp. 355–370.
[76] S. Chen, X. Tan, B. Wang, H. Lu, X. Hu, and Y. Fu, "Reverse attention-based residual network for salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 3763–3776, 2020.
[77] R. Wu, M. Feng, W. Guan, D. Wang, H. Lu, and E. Ding, "A mutual learning method for salient object detection with intertwined multi-supervision," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.
[78] Y. Xu, D. Xu, X. Hong, W. Ouyang, R. Ji, M. Xu, and G. Zhao, "Structured modeling of joint deep feature and prediction refinement for salient object detection," in *Int. Conf. Comput. Vis.*, 2019.
[79] L. Zhang, J. Zhang, Z. Lin, H. Lu, and Y. He, "Capsal: Leveraging captioning to boost semantics for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.
[80] J.-J. Liu, Q. Hou, and M.-M. Cheng, "Dynamic feature integration for simultaneous detection of salient object, edge and skeleton," *IEEE Trans. Image Process.*, vol. 29, pp. 8652–8667, 2020.
[81] Z. Wu, L. Su, and Q. Huang, "Stacked cross refinement network for edge-aware salient object detection," in *Int. Conf. Comput. Vis.*, 2019.



Jiang-Jiang Liu is currently a Ph.D. candidate with School of Computer Science, Nankai University, under Prof. Ming-Ming Cheng's supervision. His research interests include deep learning, image processing, and computer vision.

Zhi-Ang Liu received his B.S. degree from the School of Electrical Engineering and Automation, Harbin Institute of Technology in 2019. Currently, he is a master student in the College of Computer Science, Nankai University, supervised by Prof. Ming-Ming Cheng. His research interests include machine learning and computer vision.

Pai Peng received the Ph.D. degree in computer science from Zhejiang University in 2016. He is currently a senior research scientist in Youtu Lab of Tencent Technology (Shanghai) Co., Ltd. His research interests include image recognition and deep learning and has published $10+$ top-tier conference and journal papers related with visual recognition and retrieval.

Ming-Ming Cheng received his Ph.D. degree from Tsinghua University in 2012. Then he did 2 years research fellow with Prof. Philip Torr in Oxford. He is now a professor at Nankai University, leading the Media Computing Lab. His research interests include computer graphics, computer vision, and image processing. He has published $60+$ refereed research papers, with $20,000+$ Google Scholar citations. He received research awards including ACM China Rising Star Award, IBM Global SUR Award, *etc.* He is a senior member of IEEE and on the editor board of IEEE TIP.