Locate Globally, Segment Locally: A Progressive Architecture With Knowledge Review Network for Salient Object Detection

Binwei Xu, Haoran Liang*, Ronghua Liang, Peng Chen

College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China
{xubinwei, haoran, rhliang, chenpeng}@zjut.edu.cn

Abstract

Salient object location and segmentation are two different tasks in salient object detection (SOD). The former aims to globally find the most attractive objects in an image, whereas the latter can be achieved only using local regions that contain salient objects. However, previous methods mainly accomplish the two tasks simultaneously in a simple end-to-end manner, which leads to the ignorance of the differences between them. We assume that the human vision system orderly locates and segments objects, so we propose a novel progressive architecture with knowledge review network (PA-KRN) for SOD. It consists of three parts. (1) A coarse locating module (CLM) that uses body-attention label locates rough areas containing salient objects without boundary details. (2) An attention-based sampler highlights salient object regions with high resolution based on body-attention maps. (3) A fine segmenting module (FSM) finely segments salient objects. The networks applied in CLM and FSM are mainly based on our proposed knowledge review network (KRN) that utilizes the finest feature maps to reintegrate all previous layers, which can make up for the important information that is continuously diluted in the top-down path. Experiments on five benchmarks demonstrate that our single KRN can outperform state-of-the-art methods. Furthermore, our PA-KRN performs better and substantially surpasses the aforementioned methods.

Introduction

Salient object detection (SOD) has been rapidly developed recently and widely applied in many computer vision fields. As is known to all, the edge of salient objects contains rich detailed information. Hence, many methods (Zhou et al. 2020; Zhao et al. 2019; Zhao and Wu 2019; Zhang et al. 2017; Qin et al. 2019; Feng, Lu, and Ding 2019; Wu, Su, and Huang 2019b) introduce edge-related information to help identify the boundary regions of salient objects and substantially improve the accuracy of SOD. Moreover, some methods add post-processing operations (*e.g.*, CRF) (Hou et al. 2017; Li and Yu 2016; Liu, Han, and Yang 2018) to preserve fine boundary details. Although the above methods have made some progress, refining the boundaries remains a huge challenge. The first issue is that the low resolution of salient objects results in rough edge details. In particular, the segmentation of small salient objects is extremely poor. Another key issue is that the simple end-to-end structure may not be good at implementing SOD. SOD actually involves two tasks: locating and segmenting salient objects. Specifically, locating salient objects is to find the local region of salient objects from the global perspective of the whole image, whereas segmenting salient objects is to distinguish the boundary of salient objects from the local perspective of the regions that contain salient objects and their surrounding background. They are two completely different tasks and have an obvious sequence order. Moreover, the size of the images in a dataset is relatively uniform, but the size of the salient objects varies over a large range. Location is based on the relatively stable scale of global image, whereas segmentation is related to the salient objects with various scales, so segmentation clearly faces more challenges of scale variations. Many recent methods (Zhao et al. 2019; Wei, Wang, and Huang 2019; Zhao and Wu 2019; Liu et al. 2019; Feng, Lu, and Ding 2019; Pang et al. 2020) implement these two tasks simultaneously by a single end-to-end network. They don't take the differences between the two tasks into account, so breaking out of the existing results to achieve finer ones is difficult.

Unlike most of the existing end-to-end deep learning methods, the human vision system orderly locates and segments salient objects. Specifically, as shown in Fig. 1, hu-

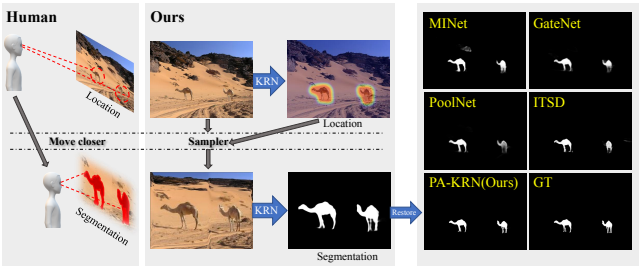

Figure 1: An illustration of our proposed PA-KRN that simulates the human biological process of globally locating salient objects then locally segmenting them. And compared with MINet (Pang et al. 2020), GateNet (Zhao et al. 2020), PoolNet (Liu et al. 2019), and ITSD (Zhou et al. 2020), our proposed method has clearer object boundaries.

*Corresponding author.

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

mans firstly find the rough areas that contain salient objects. Afterward, basing on the scale of the salient object, they adjust the distance between their eyes and the image and then focus on these areas to achieve precise segmentation. Inspired by this biological capability, we propose a progressive architecture with knowledge review network (PA-KRN). We can find that the adaptive adjustment of the distance between the eyes and an image is equivalent to indirectly increasing the resolution of salient objects while controlling the scales of salient objects within a smaller range. The PA-KRN consists of three parts. (1) A coarse locating module (CLM) that uses body-attention label locates rough areas including salient objects without boundary details. (2) An attention-based sampler highlights salient object regions with high resolution based on body-attention maps. (3) A fine segmenting module (FSM) finely segment salient objects.

U-shape networks (Lin et al. 2017; Ronneberger, Fischer, and Brox 2015) gradually integrate features of different layers from deep to shallow to obtain comprehensive information. Different layers play distinct roles, they all make indispensable contributions to SOD. Through this feature integration, the final saliency maps achieve good results. Nonetheless, some important issues still warrant attention. When the network gradually merges the next shallower feature maps, the previously learned information from deeper layers may be constantly diluted at the same time. In addition, efficiently fusing features to obtain valid information of a certain layer is difficult only by a simple fusion operation. To remedy these issues, we propose a novel knowledge review network (KRN) to efficiently acquire significant information of each layer and avoid dilution of knowledge during feature integration by introducing knowledge review (KR) module and side-out aggregation (SA) module.

To demonstrates the performance of our method, we conduct experiments on five popular benchmarks and visualize saliency maps. We implement a series of ablation studies to investigate the reliability of each module. Our main contributions are as follows:

1. A novel progressive architecture with knowledge review network (PA-KRN) is proposed to simulate the human biological process of globally locating salient objects then locally segmenting them. The PA-KRN includes coarse locating module (CLM), attention-based sampler, and fine segmenting module (FSM).
2. We design a novel knowledge review network (KRN) to avoid dilution of important information and effectively acquire significant information.
3. Extensive experiments on five popular SOD datasets demonstrate that our single KRN outperforms state-of-the-art methods. Furthermore, our PA-KRN performs better and surpassed the aforementioned models by a large margin.

Related Work

Traditional methods use mainly hand-crafted features, such as center prior (Jiang et al. 2013; Jiang and Davis 2013), texture (Yan et al. 2013), and color contrast (Cheng et al. 2014) for SOD. Basing on these low-level features, obtaining important contextual semantic information is difficult. Recently, an increasing number of SOD methods based on convolutional neural networks (CNNs) have been presented, and their performances have been gradually improved. (Hou et al. 2017) introduced short connections and combined features from different layers to generate saliency maps. Other methods mainly refine the results by improving the network structure, such as introducing attention mechanism (Chen et al. 2018; Zhao and Wu 2019; Liu, Han, and Yang 2018), iterative refining (Deng et al. 2018; Wang et al. 2019; Wei, Wang, and Huang 2019; Liu et al. 2019), and using efficient feature fusion modules (Zhang et al. 2018; Pang et al. 2020; Chen et al. 2020). These methods implement SOD in a simple end-to-end manner and don't design corresponding network according to location and segmentation.

The edge of salient objects contains rich detailed information. Hence, many approaches introduce edge-related information to help identify the boundary regions. In (Zhao and Wu 2019; Zhang et al. 2017; Qin et al. 2019; Feng, Lu, and Ding 2019), they constructed the loss function related to boundaries to emphasize the importance of edge information for SOD. (Liu et al. 2019) added a boundary branch through an additional edge dataset and achieved better performance. With the idea of the complementarity between salient object information and salient edge information as basis, (Zhao et al. 2019) utilized salient edge information to help the saliency features locate salient objects and obtain accurate object boundaries. (Zhou et al. 2020) analyzed the correlation between saliency and edge and presented an interactive two-stream decoder that explores multiple cues of the saliency and contour maps for saliency detection. Although introducing boundary information can improve results, how to get a refined saliency map remains a problem.

Methodology

In this paper, we propose a novel progressive architecture with knowledge review network (PA-KRN) that can accurately locate salient objects and improve boundary details. We describe our method in detail from two parts: overall framework and knowledge review network (KRN).

Overall Framework

Fig. 2 illustrates our overall framework, which consists of three parts. 1) The CLM is a network that locates salient objects. We design a body-attention map as the label, which mainly concentrates on rough regions, including salient objects, and ignores edge details. Without the disturbance of pixels around the edges of salient objects, the body-attention map can guide the model to obtain better representations. 2) The attention-based sampler highlights salient object regions with high resolution based on the body-attention map. 3) The FSM, similar to other networks for SOD, is a network that can individually achieve the task of SOD. The difference is that the input images of FSM are pre-processed with higher resolution on salient objects and smaller scale variations between salient objects. The networks applied in CLM and FSM are mainly based on knowledge review network

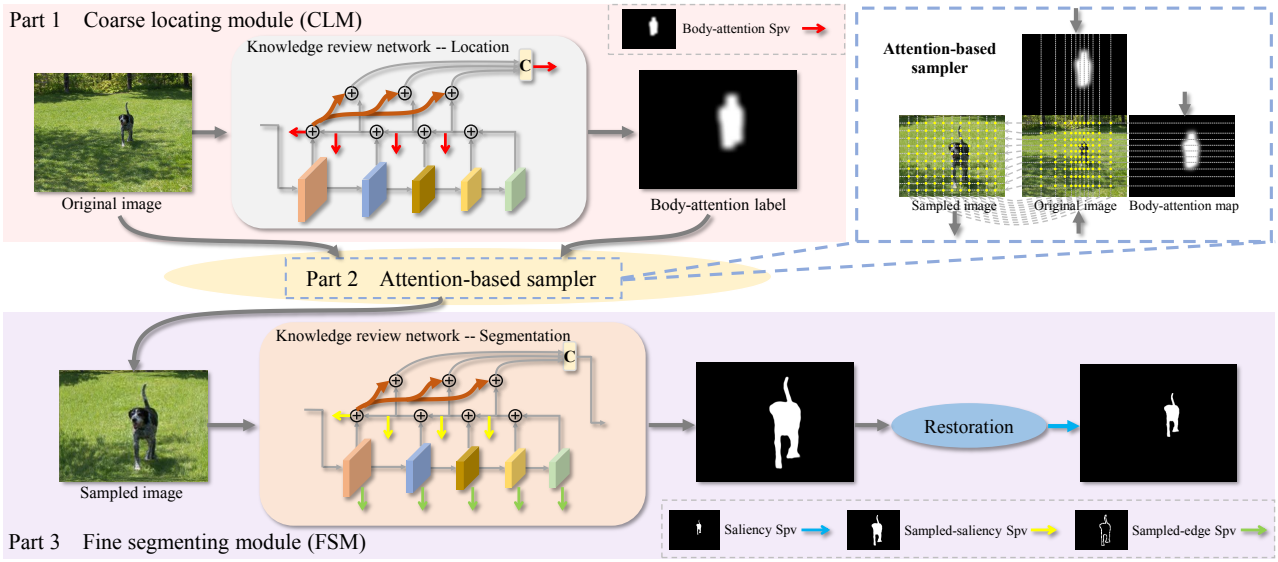

Figure 2: Overview of our proposed method. It consists of three parts: coarse locating module (CLM) that locates salient objects, attention-based sampler that highlights salient object regions with high resolution, and fine segmenting module (FSM) that completes the segmentation task. The network structures of CLM and FSM are based on knowledge review network (KRN).

(KRN). The whole system can be trained in an end-to-end manner. To get finer results, we first train CLM and FSM sequentially and then combine the three parts to fine-tune.

Coarse Locating Module The goal of CLM is to find the precise location of salient objects that will lay a robust foundation for the subsequent accurate segmentation. The problem is how to convey the location information of salient objects by the form of images. The original label retains detailed location information. However, (Wei et al. 2020) observed that pixels close to the edge are likely to be mispredicted due to unbalanced distribution. Hence, the original label cannot be directly used to avoid incompleteness of objects. Inspired by the task of fixation prediction (Huang et al. 2015; Kümmerer, Wallis, and Bethge 2016), we propose a body-attention label, which concentrates mainly on location information, to guide the network.

To fully contain the salient object and smooth the edge, we enlarge the label by a binary dilation operation with a kernel size of $K \times K$. Moreover, to facilitate the model to learn the object location distribution and retain the background information around for segmentation, we simply apply Gaussian blur with the sigma of 8 and the same kernel size as dilation operation to generate body-attention label. Fig. 3 shows examples of body-attention maps. We can observe that the body-attention label can not only smooth the complex edge (row 3) but also make the thin parts of the salient object easy to detect by expanding them (row 1). It also fills the region of interference objects inside the salient object to strengthen the integrity of the object region (row 2).

Attention-based Sampler After obtaining body-attention maps, we use them to increase the resolution of the regions related to salient objects in an image, which can magnify the details of the salient object. And salient objects are magnified closer to the image size, which will narrow the scale gap between salient objects of a dataset. We introduce attention-based sampler that is proposed in the fine-grained classification task (Zheng et al. 2019) to accomplish our goal. As shown in Fig. 2, the main idea is to sample the pixels of the original image according to the attention value of its body-attention map. To be specific, areas with high attention value are sampled more intensively.

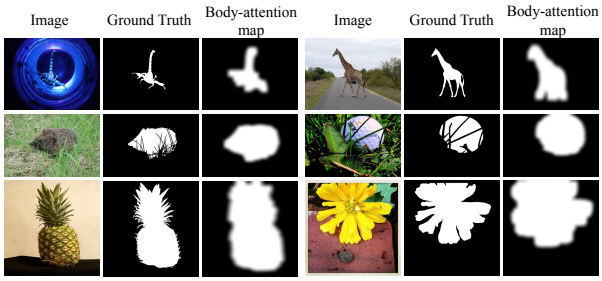

Figure 3: Examples of body-attention label.

Fine Segmenting Module Similar to other SOD networks in terms of tasks and functions, FSM needs to complete the tasks of salient object location and segmentation. The biggest difference is that the input images of FSM are preprocessed. Salient objects in these images have more uniform scales and higher resolution, so they can effectively help refine the boundaries of salient objects. Additionally, in processed images, salient objects with high resolution are located around the central region, and the background is compressed, thus the difficulty of locating salient objects is greatly reduced. At the last step, the output is restored to the

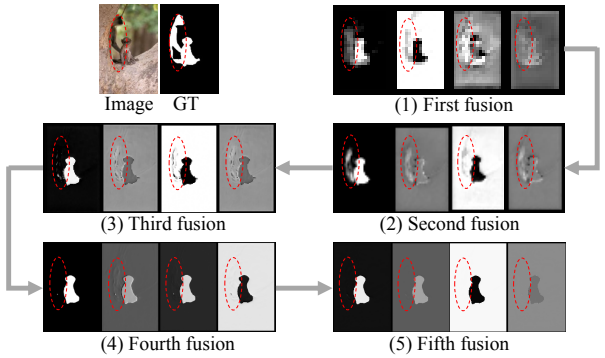

Figure 4: An example of information dilution. We visualize four representative feature maps that are sampled after each fusion of the top-down path in FPNs. Apparently, the salient object in the red circle gradually disappear with the continuous fusion process.

saliency map corresponding to the original image.

Knowledge Review Network

Classic U-shape networks like feature pyramid networks (FPNs) (Lin et al. 2017) gradually integrate features layer by layer through the top-down transmission to obtain comprehensive information, but they still have shortcomings. As shown in Fig. 4, we extract four feature maps of an image after each fusion of the top-down path in FPNs. The salient object in the red circle does not further restore the details but gradually disappear with the continuous fusion process, which means the key information of high-level features will be gradually diluted. In addition, feature fusion is inadequate to efficiently distinguish and obtain useful information. To remedy the above problems, we propose knowledge review network (KRN), as shown in Fig. 5. It is based on the U-shape FPNs with the pretrained ResNet-50 as the backbone, which is a bottom-up and top-down encoder-decoder that can fully combine multi-scale features to obtain rich semantic information. We design KR module to review unlearned and diluted information by recombining the finest maps with features of each layer and add SA module to improve the efficiency of feature fusion.

Note that KRN is employed in CLM and FSM, but differences exist between them. Like a general SOD network, FSM needs to accurately distinguish salient objects from the pixel level. Thus, we add the intermediate edge supervisions to guide the features provided by the encoding process to have clear boundaries. Single KRN with edge supervisions can finely complete SOD alone, so we train it alone. Then we use it to evaluate its each module and compare it with other methods to evaluate its performance. We denote it as SGL-KRN.

Knowledge Review Module We present the KR module to recombine the finest feature maps with features of each layer. Though the review process, valid information can be captured again to dramatically improve the utilization of these features. Specifically, as shown in Fig. 5, there are five

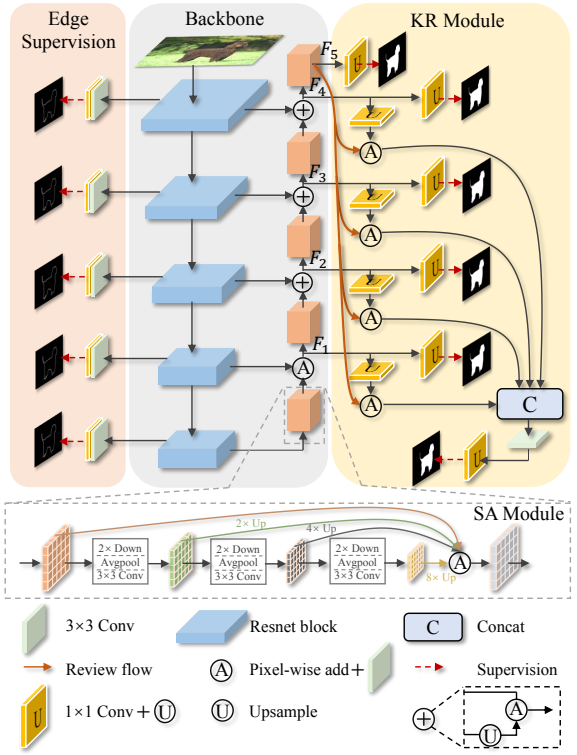

Figure 5: Detailed illustration of our knowledge review network. It adds KR module, SA module, and the intermediate edge supervisions on the basis of FPNs. Intermediate edge supervision exists in CLM but not in FSM.

groups of feature maps after fusions, which are respectively named F_1, F_2, F_3, F_4, and F_5 according to the resolution from low to high. Different groups of feature maps retain varying levels of details and semantic information. But they have distinct sizes and numbers. We first squeeze the channels of F_1, F_2, F_3, and F_4 to be the same as F_5 by a 1×1 convolution. Following, we resize these feature maps to the same size as F_5 by upsampling operation. They are then fused with F_5 to supplement the diluted and undiscovered important information by a pixel-wise add operation and a 3×3 convolution. To avoid introducing interference information that is produced by the large difference between the finest feature maps and rough top-layers feature maps, we add the intermediate supervision to guide all feature maps to retain only the helpful information related to salient objects. Next, these four groups of fused feature maps are integrated by a concatenation operation and a 3×3 convolution. The final saliency map will be generated by a 1×1 convolution and an upsampling operation.

Side-out Aggregation Module The re-fusion of features in KR module can make up for the missing or unfused helpful information, but obtaining useful information as efficiently as possible during feature integration can further improve the availability of the features. To this end, we add a simple SA module during feature integration to improve the efficiency of feature fusion. As shown in Fig. 5, sim-

ilar to FPNs, feature maps of multiple scales are obtained by multiple down-sampling, average pooling, and a 3×3 convolution. Afterward, we simply merge all feature maps, followed by a 3×3 convolution filter. By combining multi-scale features, more comprehensive information can be extracted from different scale spaces to avoid the omission of important information. In addition, SA module can further enhance the receptive field of the whole network.

Loss Function

Similar to fixation prediction, the goal of CLM is to predict the distribution of salient objects rather than the exact value of each pixel. Sen (Jia and Bruce 2020) designed a loss function that consists of three saliency evaluation metrics, namely linear correlation coefficient (CC), Kullback-Leibler divergence (KLD), and normalized scanpath saliency (NSS). We further modify NSS in the combined loss function to meet our task. The loss function of body-attention supervision is $l_b = NSS' + CC' + KLD$, where NSS' is the variant of NSS, and CC' is the variant of CC. We denote predicted saliency maps as P and body-attention maps as Q. In addition, we extract these pixels with values of 255 in Q to construct a new ground truth as F, which is the region where there is a fairly high probability of salient objects. NSS is used to measure the average normalized values of P at the eye fixation points in fixation prediction F (Peters et al. 2005), which emphasizes the importance of these fixation points. In the task of locating salient object, what we need to focus on is the pixels with high value in F. NSS' in our method is shown as:

$$NSS'(P,F) = \frac{1}{N}\sum_{i}\left(\frac{F-\mu(F)}{\sigma(F)} - \frac{P-\mu(P)}{\sigma(P)}\right) \times F_i \quad (1)$$

where i indexes the i^{th} pixel, N is the number of high value pixels in F, $\mu(\cdot)$ and $\sigma(\cdot)$ represent the mean and standard deviation of the input, respectively. The total loss of CLM is:

$$L_{clm} = \lambda_1 l_b + \lambda_2 \sum_{i=1}^{5} l_b^i \quad (2)$$

where l_b^i is the loss of the i^{th} intermediate body-attention supervision. λ_1 and λ_2 is set to 2 and 1. Fig. 2 shows all supervisions.

The loss function of saliency supervision is $l_s = l_{bce} + l_{iou}$, where l_{bce} and l_{iou} are BCE loss (De Boer et al. 2005) and IoU loss (Máttyus, Luo, and Urtasun 2017). The loss function of sampled saliency supervision l_{sa} is the same, except that the ground truth is sampled. l_e is the loss function of edge supervision, and we use BCE loss to construct it. The total loss of FSM is:

$$L_{fsm} = \lambda_3 l_s + \lambda_4 \sum_{i=1}^{5} l_{sa}^i + \lambda_5 \sum_{i=1}^{5} l_e^i \quad (3)$$

where l_{sa}^i is the loss of the i^{th} intermediate sampled-saliency supervision, and l_e^i is the loss of the i^{th} intermediate sampled-edge supervision. λ_3, λ_4 and λ_5 is set to 2, 1 and 1 respectively. In jointing training, the total loss L is expressed as:

$$L = L_{clm} + L_{fsm} \quad (4)$$

Experiments

Datasets and Evaluation Metrics

To validate the performance of our proposed method, we conduct experiments on five popular benchmark datasets, namely, ECSSD (Yan et al. 2013), DUT-OMRON (Yang et al. 2013), HKU-IS (Li and Yu 2015), DUTS (Wang et al. 2017), and PASCAL-S (Li et al. 2014). DUTS is the largest available dataset among them, which contains 10553 and 5019 images for training (DUT-TR) and testing (DUT-TE) respectively. Thus, we train the model on DUT-TR and test on the other five datasets.

To quantitatively evaluate the performance of our methods and existing state-of-the-art approaches, we adopt four widely used metrics, which are the precision-recall curves (PR curves), F-measure (Achanta et al. 2009) and curves, mean absolute error (MAE), and E-measure (Fan et al. 2018). We use the max f-measure over all thresholds from 0 to 255, denoted as F_{max}.

Implementation Details

We use horizontal flip, randomly rotate, and multi-scale input images for data augmentation and adopt Adam optimizer (Kingma and Ba 2014) with a weight decay of 5e-4 and learning rate of 5e-5 which is divided by 10 after 15 epochs to train CLM and FSM. These modules are trained for 24 epochs. The backbone parameters are initialized from the ResNet-50 pretrained on the ImageNet dataset (Krizhevsky, Sutskever, and Hinton 2012). For joint training, learning rate is set to 5e-6 which is divided by 10 after 9 epochs, and the total number of epochs is set to 15. In both training and testing, we keep the sizes of the input unchanged as done in (Liu et al. 2019) and do not use any post-processing.

Comparison with the State-of-the-arts

We compare our model with 14 state-of-the-art methods. For fair comparison, the results of these 14 methods are directly provided by the author or by their original trained model and we test them with the same evaluation codes.

Quantitative Comparison As shown in Tab. 1, we compare SGL-KRN and PA-KRN with other methods in terms of F_{max}, F_{avg}, E-measure, and MAE. The SGL-KRN has shown good performance and significantly outperforms other methods, demonstrating the effectiveness of our proposed KRN. Specifically, SGL-KRN outperforms previous methods by a large margin on DUT-OMRON, HKU-IS, DUTS-TE under different measurements. Although it does not perform best compared with other methods on ECSSD, it is also very competitive and close to the best one. Furthermore, PA-KRN achieves better results and obviously surpasses other methods on all datasets. Fig. 6 demonstrates the standard precision-recall curves and F-measure curves. SGL-KRN achieves the best results compared with other methods on the DUTS-TE, DUTS-OMRON, PASCAL-S, and HKU-IS datasets and is very competitive on ECSSD. These results show that our proposed KRN has a good capability to produce high-quality saliency maps. Furthermore,

	ECSSD			DUT-OMRON			HKU-IS			DUTS-TE			PASCAL-S		
	F_{max}	E_m	MAE	F_{max}	E_m	MAE	F_{max}	E_m	MAE	F_{max}	E_m	MAE	F_{max}	E_m	MAE
RAS (Chen et al. 2018)	0.921	0.914	0.056	0.786	0.846	0.062	0.913	0.929	0.045	0.831	0.861	0.059	0.829	0.829	0.101
DGRL (Wang et al. 2018)	0.925	0.917	0.043	0.779	0.843	0.063	0.913	0.941	0.037	0.828	0.897	0.050	0.848	0.834	0.073
PiCANet (Liu, Han, and Yang 2018)	0.940	0.919	0.035	0.804	0.859	0.052	0.927	0.945	0.031	0.867	0.894	0.040	0.859	0.844	**0.063**
MLMSNet (Wu et al. 2019)	0.928	0.914	0.045	0.774	0.837	0.064	0.921	0.937	0.039	0.852	0.860	0.049	0.850	0.836	0.073
AFNet (Feng, Lu, and Ding 2019)	0.935	0.918	0.042	0.797	0.853	0.057	0.923	0.942	0.036	0.863	0.879	0.046	0.858	0.845	0.069
PS (Wang et al. 2019)	0.938	0.922	0.041	0.812	0.854	0.061	0.922	0.942	0.038	0.855	0.879	0.048	0.855	0.850	0.070
CPD (Wu, Su, and Huang 2019a)	0.939	**0.924**	0.037	0.797	0.866	0.056	0.925	0.944	0.034	0.865	0.887	0.043	0.859	0.849	0.070
BASNet (Qin et al. 2019)	0.942	0.921	0.037	0.805	0.869	0.056	0.928	0.946	0.032	0.860	0.884	0.048	0.854	0.846	0.075
PoolNet (Liu et al. 2019)	0.944	**0.924**	0.039	0.808	0.863	0.056	0.932	0.948	0.033	0.880	0.889	0.040	0.863	0.848	0.074
EGNet (Zhao et al. 2019)	0.947	**0.927**	0.037	0.815	0.867	0.053	0.935	0.950	0.031	0.889	0.891	0.039	0.865	0.848	0.073
ITSD (Zhou et al. 2020)	0.947	**0.927**	0.034	0.821	0.863	0.061	0.934	0.952	0.031	0.883	0.895	0.041	0.870	0.850	0.065
GCPANet (Chen et al. 2020)	**0.948**	0.920	0.035	0.812	0.860	0.056	0.938	0.949	0.031	0.888	0.891	0.038	0.869	0.847	**0.061**
GateNet (Zhao et al. 2020)	0.945	**0.924**	0.040	0.818	0.862	0.055	0.933	0.949	0.033	0.888	0.889	0.040	0.869	0.851	0.067
MINet (Pang et al. 2020)	0.947	**0.927**	**0.033**	0.810	0.865	0.055	0.935	0.953	0.029	0.884	0.898	0.037	0.867	0.851	**0.063**
SGL-KRN (Ours)	0.946	**0.927**	0.036	**0.827**	**0.883**	**0.049**	**0.939**	**0.954**	**0.028**	**0.898**	**0.913**	**0.034**	**0.872**	**0.859**	0.067
PA-KRN (Ours)	**0.953**	**0.924**	**0.032**	**0.834**	**0.885**	**0.050**	**0.943**	**0.955**	**0.027**	**0.907**	**0.916**	**0.033**	**0.873**	**0.857**	0.066

Table 1: Quantitative comparisons with 14 state-of-the-art methods on five datasets with max F-measure score, MAE and E-measure. The best two results are marked in bold. SGL-KRN: single KRN trained by DUT-TR.

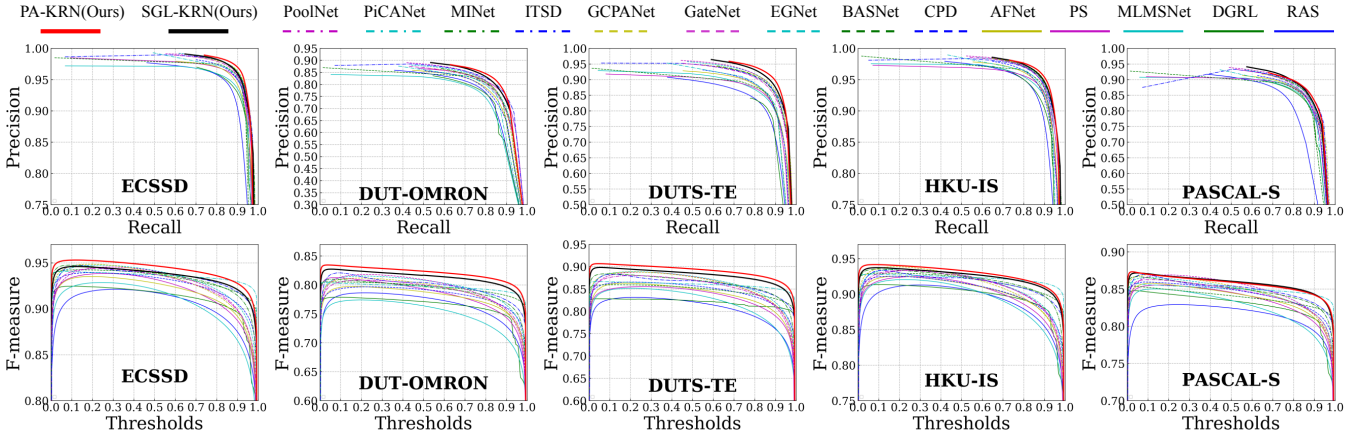

Figure 6: Performance comparison with 14 state-of-the-art algorithms on 5 datasets. The first row shows PR curves and the second row shows F-measure curves.

the curve of PA-KRN is obviously lying above others, which demonstrates PA-KRN is absolutely effective and robust.

Visual Comparison To further verify the advantages of our method, we provide visual examples of the proposed methods and other state-of-the-art approaches in Fig. 7. Our methods can effectively highlight salient objects in various challenging scenarios, including objects with similar appearances to backgrounds (row 1), complex backgrounds (row 2 and 3), large foregrounds (row 2), tiny objects (row 4 and 6), and slender object (row 5). To be specific, compared with other approaches, we can see that both SGL-KRN and PA-KRN accurately find salient objects (row 3, and 4), which shows the robustness of our and KRN. In row 1 and 2, our PA-KRN can accurately distinguish salient objects but SGL-KRN cannot, which shows the effectiveness of body-attention label in CLM and the progressive architecture. In addition, we can find that PA-KRN has a distinct advantage in detail processing, and can clearly distinguish the boundaries of salient objects, even for challenging small objects and slender objects (rows 4, 5, and 6). These outcomes demonstrate the superiority of our method.

Ablation Study

Key Components in KRN: To evaluate the effectiveness of different modules in KRN, we conduct a series of ablation experiments based on FPNs baseline on DUTS-TE and DUTS-OMRON, as shown in Tab. 2. For fair comparison, except for combinations of different components, other configurations are the same. As we can see, whether or not the SA module is added, when the KR module is introduced, the results of our model are significantly improved. It demonstrates that our proposed KR module is effective and necessary for good performance. In addition, we can observe that the SA module can effectively boost the performance, which means that it can bring substantial benefit to SOD. When in-

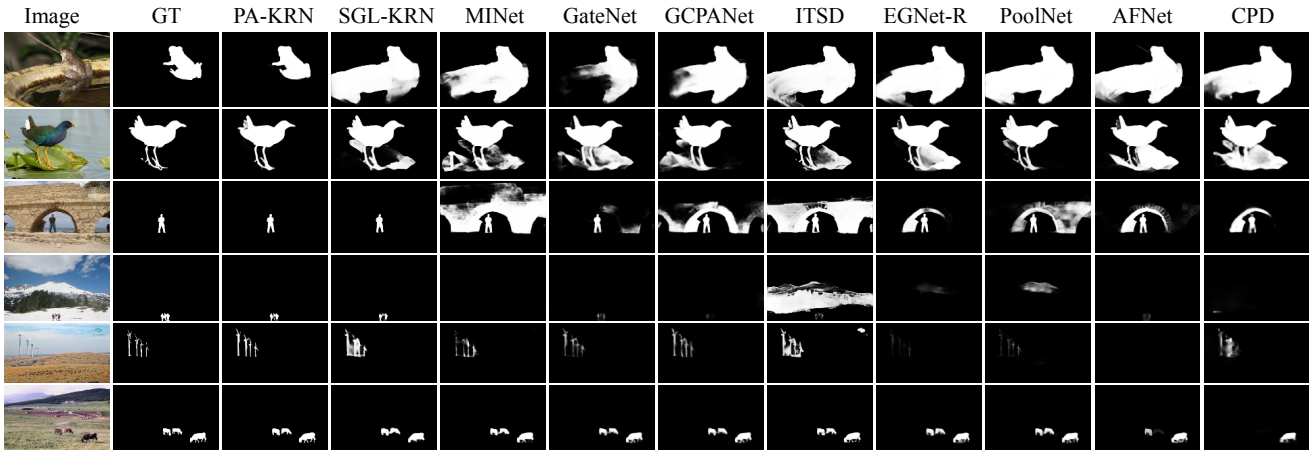

Figure 7: Visual comparisons of different methods. Each row shows saliency maps of one image. Each column represents one algorithms. Apparently, our method can more accurately find salient objects and more clearly distinguish the boundaries of salient objects than other state-of-the-art methods.

Baseline	KR	SA	Edge	DUTS-TE			DUT-OMRON		
				F_{max}	E_m	MAE	F_{max}	E_m	MAE
√				0.871	0.893	0.043	0.799	0.864	0.058
√	√			0.884	0.903	0.038	0.818	0.873	0.055
√		√		0.887	0.902	0.039	0.822	0.874	0.053
√	√	√		0.895	0.908	0.035	0.827	0.881	0.050
√	√	√	√	0.898	0.913	0.034	0.827	0.883	0.049

Table 2: Ablation analysis for the key components (*i.e.*, KR module, SA module, and edge supervision) in KRN on two challenging datasets. ResNet-based FPNs are used as the baseline.

termediate edge supervisions are combined, we can obtain slightly better results. Simply adding intermediate edge supervision exerts a certain effect.

Effectiveness of Progressive Architecture: To further investigate the effectiveness of our proposed progressive architecture that globally locates and locally segments, we compare the results for the different networks (*i.e.*, U-Net (Ronneberger, Fischer, and Brox 2015), CPD, PoolNet, and our KRN) embedded in our proposed progressive architecture (PA) or not, as shown in Tab. 3. Specifically, these networks integrate into PA by replacing the KRN in PA-KRN and keep their original loss function unchanged. For fair comparison, configurations are the same for each method. We can see that all methods perform significantly better, which proves that our proposed progressive architecture is reliable and practical.

Kernel Size in Body-attention Label: There is a vital hyperparameter K to be determined. K is used as the kernel size of the dilation operation and the Gaussian filter when generating the body-attention label. We set K to fixed values of 15,25, and 35, and adaptive values that are $\frac{1}{15}$, $\frac{1}{10}$, and $\frac{1}{5}$ of the shorter side length respectively. As shown in Tab. 4, we implement experiments on DUT-TE and the model performed best when K is set to a fixed value of 25.

Method	PA	DUTS-TE			DUT-OMRON		
		F_{max}	E_m	MAE	F_{max}	E_m	MAE
U-Net	×	0.736	0.791	0.086	0.659	0.764	0.101
	√	**0.751**	**0.802**	**0.083**	**0.675**	**0.773**	**0.097**
CPD	×	0.865	0.887	**0.043**	0.798	0.862	0.057
	√	**0.874**	**0.892**	**0.043**	**0.807**	**0.868**	**0.056**
PoolNet	×	0.881	0.889	0.040	0.807	0.862	0.056
	√	**0.894**	**0.906**	**0.038**	**0.824**	**0.879**	**0.054**
Ours	×	0.898	0.913	0.034	0.827	0.883	**0.049**
	√	**0.907**	**0.916**	**0.033**	**0.834**	**0.885**	0.050

Table 3: Quantitative comparison of different methods with or without our progressive architecture (PA).

Fixed K	DUTS-TE			Adaptive K	DUTS-TE		
	F_{max}	E_m	MAE		F_{max}	E_m	MAE
15	0.904	0.914	**0.033**	$\frac{1}{15}$	0.904	**0.916**	**0.033**
25	**0.907**	**0.916**	**0.033**	$\frac{1}{10}$	0.903	0.012	0.034
35	0.902	0.912	0.034	$\frac{1}{5}$	0.900	0.908	0.035

Table 4: Ablation analysis for different kernel sizes of the dilation operation and the Gaussian filter in body-attention label.

Conclusion

We have presented a novel progressive architecture with knowledge review network for SOD, which simulates the biological capability of humans to globally locate and locally segment salient objects sequentially. In addition, to improve the network performance in the framework, we have proposed a novel knowledge review network to make full use of the information of each layer by recombining finest feature maps with those of previous layers. Extensive experiments well demonstrate that the proposed method outperforms state-of-the-art methods under different benchmarks.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2020YFB1707700) and the National Natural Science Foundation of China (61702457, 62036009, 61871350, U1909203).

References

Achanta, R.; Hemami, S.; Estrada, F.; and Susstrunk, S. 2009. Frequency-tuned salient region detection. In *2009 IEEE conference on computer vision and pattern recognition*, 1597–1604. IEEE.

Chen, S.; Tan, X.; Wang, B.; and Hu, X. 2018. Reverse attention for salient object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 234–250.

Chen, Z.; Xu, Q.; Cong, R.; and Huang, Q. 2020. Global context-aware progressive aggregation network for salient object detection. *arXiv preprint arXiv:2003.00651* .

Cheng, M.-M.; Mitra, N. J.; Huang, X.; Torr, P. H.; and Hu, S.-M. 2014. Global contrast based salient region detection. *IEEE transactions on pattern analysis and machine intelligence* 37(3): 569–582.

De Boer, P.-T.; Kroese, D. P.; Mannor, S.; and Rubinstein, R. Y. 2005. A tutorial on the cross-entropy method. *Annals of operations research* 134(1): 19–67.

Deng, Z.; Hu, X.; Zhu, L.; Xu, X.; Qin, J.; Han, G.; and Heng, P.-A. 2018. R3net: Recurrent residual refinement network for saliency detection. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 684–690. AAAI Press.

Fan, D.-P.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M.-M.; and Borji, A. 2018. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421* .

Feng, M.; Lu, H.; and Ding, E. 2019. Attentive feedback network for boundary-aware salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1623–1632.

Hou, Q.; Cheng, M.-M.; Hu, X.; Borji, A.; Tu, Z.; and Torr, P. H. 2017. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3203–3212.

Huang, X.; Shen, C.; Boix, X.; and Zhao, Q. 2015. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 262–270.

Jia, S.; and Bruce, N. D. 2020. Eml-net: An expandable multi-layer network for saliency prediction. *Image and Vision Computing* 103887.

Jiang, H.; Wang, J.; Yuan, Z.; Wu, Y.; Zheng, N.; and Li, S. 2013. Salient object detection: A discriminative regional feature integration approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2083–2090.

Jiang, Z.; and Davis, L. S. 2013. Submodular salient region detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2043–2050.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Kümmerer, M.; Wallis, T. S.; and Bethge, M. 2016. DeepGaze II: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563* .

Li, G.; and Yu, Y. 2015. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5455–5463.

Li, G.; and Yu, Y. 2016. Deep contrast learning for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 478–487.

Li, Y.; Hou, X.; Koch, C.; Rehg, J. M.; and Yuille, A. L. 2014. The secrets of salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 280–287.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Liu, J.-J.; Hou, Q.; Cheng, M.-M.; Feng, J.; and Jiang, J. 2019. A simple pooling-based design for real-time salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3917–3926.

Liu, N.; Han, J.; and Yang, M.-H. 2018. Picanet: Learning pixel-wise contextual attention for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3089–3098.

Máttyus, G.; Luo, W.; and Urtasun, R. 2017. Deeproadmapper: Extracting road topology from aerial images. In *Proceedings of the IEEE International Conference on Computer Vision*, 3438–3446.

Pang, Y.; Zhao, X.; Zhang, L.; and Lu, H. 2020. Multi-Scale Interactive Network for Salient Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9413–9422.

Peters, R. J.; Iyer, A.; Koch, C.; and Itti, L. 2005. Components of bottom-up gaze allocation in natural scenes. *Journal of Vision* 5(8): 692–692.

Qin, X.; Zhang, Z.; Huang, C.; Gao, C.; Dehghan, M.; and Jagersand, M. 2019. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7479–7489.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Wang, L.; Lu, H.; Wang, Y.; Feng, M.; Wang, D.; Yin, B.; and Ruan, X. 2017. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 136–145.

Wang, T.; Zhang, L.; Wang, S.; Lu, H.; Yang, G.; Ruan, X.; and Borji, A. 2018. Detect globally, refine locally: A novel approach to saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3127–3135.

Wang, W.; Shen, J.; Cheng, M.-M.; and Shao, L. 2019. An iterative and cooperative top-down and bottom-up inference network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5968–5977.

Wei, J.; Wang, S.; and Huang, Q. 2019. F3Net: Fusion, feedback and focus for salient object detection. *arXiv preprint arXiv:1911.11445* .

Wei, J.; Wang, S.; Wu, Z.; Su, C.; Huang, Q.; and Tian, Q. 2020. Label Decoupling Framework for Salient Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13025–13034.

Wu, R.; Feng, M.; Guan, W.; Wang, D.; Lu, H.; and Ding, E. 2019. A mutual learning method for salient object detection with intertwined multi-supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8150–8159.

Wu, Z.; Su, L.; and Huang, Q. 2019a. Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3907–3916.

Wu, Z.; Su, L.; and Huang, Q. 2019b. Stacked cross refinement network for edge-aware salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 7264–7273.

Yan, Q.; Xu, L.; Shi, J.; and Jia, J. 2013. Hierarchical saliency detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1155–1162.

Yang, C.; Zhang, L.; Lu, H.; Ruan, X.; and Yang, M.-H. 2013. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3166–3173.

Zhang, L.; Dai, J.; Lu, H.; He, Y.; and Wang, G. 2018. A bi-directional message passing model for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1741–1750.

Zhang, P.; Wang, D.; Lu, H.; Wang, H.; and Ruan, X. 2017. Amulet: Aggregating multi-level convolutional features for salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 202–211.

Zhao, J.-X.; Liu, J.-J.; Fan, D.-P.; Cao, Y.; Yang, J.; and Cheng, M.-M. 2019. EGNet: Edge guidance network for salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 8779–8788.

Zhao, T.; and Wu, X. 2019. Pyramid feature attention network for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3085–3094.

Zhao, X.; Pang, Y.; Zhang, L.; Lu, H.; and Zhang, L. 2020. Suppress and balance: A simple gated network for salient object detection. *arXiv preprint arXiv:2007.08074* .

Zheng, H.; Fu, J.; Zha, Z.-J.; and Luo, J. 2019. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5012–5021.

Zhou, H.; Xie, X.; Lai, J.-H.; Chen, Z.; and Yang, L. 2020. Interactive Two-Stream Decoder for Accurate and Fast Saliency Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9141–9150.