KD-SCFNet: Towards More Accurate and Efficient Salient Object Detection via Knowledge Distillation

Jin Zhang[1], Qiuwei Liang[2], and Yanjiao Shi[1⋆]

[1] Shanghai Institute of Technology, China
[2] Wenzhou Medical University, China

Abstract. Most existing salient object detection (SOD) models are difficult to apply due to the complex and huge model structures. Although some lightweight models are proposed, the accuracy is barely satisfactory. In this paper, we design a novel semantics-guided contextual fusion network (SCFNet) that focuses on the interactive fusion of multi-level features for accurate and efficient salient object detection. Furthermore, we apply knowledge distillation to SOD task and provide a sizeable dataset KD-SOD80K. In detail, we transfer the rich knowledge from a seasoned teacher to the untrained SCFNet through unlabeled images, enabling SCFNet to learn a strong generalization ability to detect salient objects more accurately. The knowledge distillation based SCFNet (KD-SCFNet) achieves comparable accuracy to the state-of-the-art heavyweight methods with less than 1M parameters and 174 FPS real-time detection speed. Extensive experiments demonstrate the robustness and effectiveness of the proposed distillation method and SOD framework. Code and data: https://github.com/zhangjinCV/KD-SCFNet.

Keywords: Salient object detection; Knowledge distillation; Lightweight model

1 Introduction

Salient object detection (SOD) is a basic task of computer vision. It aims to detect the most attractive area in an image by imitating the attention mechanism of human vision. As an efficient preprocessing technique, SOD is widely served for many computer vision tasks, such as image translation [12], object tracking [47,48], semantic segmentation [44,32], image retrieval [36] and so on.

In recent years, owing to the powerful capacity of extracting high-level semantic information, convolutional neural network (CNN) has been widely applied to SOD task. CNN-based SOD methods [30,40,17,23,28,38,21,49,18,22,50,19] have achieved remarkable performance and promoted SOD to realize milestone development. However, we argue that there are still some core challenges in SOD to

[⋆] Corresponding author: shiyanjiao616@163.com overcome. First, most existing SOD methods use robust encoder such as VGG-16 [33] or ResNet-50 [8] to obtain multi-scale high- and low-level features, and then construct complex feature processing modules to increase the depth or width of the network for better performance, which leads to problems such as heavy models, slow detection speed, and inconvenient to implement. For example, MINet [28] with the ResNet-50 backbone has 162M parameters and needs heavy computation overhead (105.34G FLOPs) to execute. Some works consider a compromise between speed and accuracy, and build models for rapid detection, such as U^{2}Net [30], HVPNet [22] and SAMNet [23]. Although these models improve the detection speed by reducing the complexity, the loss of accuracy is much more than the gain in speed (see Fig. 1). Second, fully supervised training with pixel-level annotated images is currently the most mainstream method. However, hand-made pixel-level annotated images require a large amount of time and effort, and the effect of the model is also limited by the number of training images, which leads to a bottleneck in the performance of existing SOD methods. Third, high-level features bring coarse resolution that may dilute the boundary of salient object. Some works [50,18,26,38,17] introduce boundary map or increase the penalty of the object's boundary to explicitly learn salient object boundary. Nevertheless, all these methods use a fixed boundary thickness throughout training. Actually, thinner boundary map promotes the model to outline the object accurately but leads to large boundary loss fluctuations at the early training stage due to the uncertain salient object and blurry boundary. Thicker boundary map is more advantageous for perceiving object's position [38] and smoothing the fluctuation, but not confident of assisting to generate object with clear boundary at the late training stage. Therefore, designing a mechanism to capture the boundary difference in different training stages to reduce the impact of boundary bias is essential. Based on the above observations, it is extremely significant and challenging to establish a salient object detection framework that takes both accuracy and efficiency into account.

Facing the above challenges, we propose a lighter, faster and wiser network, for salient object detection. Firstly, to build a lightweight model, we choose MobileNet V3 [10] as the backbone network taking the parameters and the feature extraction ability into account. It is worth noting that the U-shaped encoder-decoder architecture causes deep features to be gradually diluted as they are transferred to lower layers [18]. Some works [18,4] improved this problem by feeding deep features into each fusion layer separately. Liu et al. [18] performed pixel-wise addition of adjacent layer features and deepest features to obtain more contextual information, while Chen et al. [4] adopted pixel-wise multiplication to fuse features better. Although these methods can achieve good results, they require additional modules to optimize the fused features due to the rude fusion. Intuitively, the deep features have higher-level semantic information, which can better guide the shallow features to locate the salient objects [26], and the low-level contextual features with detail information are more beneficial to produce objects with fine boundaries [50]. Thus, we design a semantics-guided contextual fusion module (SCFM) for effectively integrating the complementary

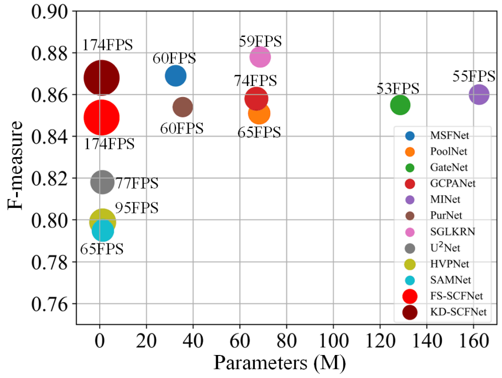

Fig. 1. The comparison between our methods (FS-SCFNet, KD-SCFNet) and existing methods in terms of parameters, accuracy and speed. The metrics F-measure are calculated in the DUTS-TE dataset [37]. Different colors indicate different methods, and the size of the circle represents the speed of detection.

information of low and deep-level features. Precisely, SCFM captures salient objects more accurately by fusing the global deep-level feature into the low-level contextual features, and then generates salient objects with fine boundaries by fusing the low-level contextual features. Thus, we propose a novel lightweight SOD framework named semantics-guided contextual fusion network (SCFNet), which simply consists of the encoder MobileNet V3 and the decoder composed of SCFMs. Secondly, to break the limitation of pixel-level annotated images and obtain better-performing models, we introduce knowledge distillation to SOD task and propose a knowledge distillation based SCFNet (KD-SCFNet). Knowledge distillation is formally proposed by Hinton et al. [9], which passes dark knowledge from complicated teachers to compact students, enabling students to maintain strong generalization as teachers. In this paper, a teacher model with more representing ability than the lightweight network is constructed and trained to transfer dark knowledge to the untrained student model. Specifically, the teacher is trained on a commonly used SOD training dataset, and the predicted results on natural images are regarded as weak labels to supervise the untrained SCFNet. Due to the knowledge distilled and transferred from the teacher model, KD-SCFNet will learn a stronge generalization ability to detect accurate salient objects. It is worth mentioning that, to distinguish from the knowledge distillation based SCFNet, the SCFNet trained in a fully supervised manner is named as FS-SCFNet. Finally, we design a structure polishing (SP) loss to improve sample imbalance issues in saliency and boundary maps. More importantly, we propose a progressive boundary supervision method (PBSM) for SP loss, which uses thicker boundary map in the early training stage to locate the salient object better, and gradually refine the boundary map with the process of training to help generate salient object with exquisite boundary. In short, our main contributions can be summarized as follows:

- We design a semantics-guided contextual fusion module (SCFM) that effectively utilizes the high-level semantic information of deep features and the detailed information of low-level contextual features, and build a lightweight salient object detection model based on SCFM.
- We introduce knowledge distillation to SOD task. The knowledge distilled from a pre-trained teacher is transferred to the student, making it more outstanding than that trained in a fully supervised way. As we know, we are the first to apply knowledge distillation to SOD task. Furthermore, we provide a large unlabeled dataset KD-SOD80K for salient object detection.
- We propose a structure polishing loss and a progressive boundary supervision method, which draw more attention to boundary details and assist the network to alleviate the adverse effects due to boundary deviations during different training stages.
- We conduct extensive experiments on 6 datasets and compare the proposed lightweight model with 15 existing state-of-the-art SOD methods. The results of the distillation experiment and the comparison experiment demonstrate the effectiveness and superiority of the proposed method.

2 Related Work

2.1 Salient Object Detection

Traditional SOD methods [11,46,5,42,29,43] mainly rely on hand-made low-level features [11,5,42] and heuristic clues [46,29,43]. With the development of deep learning technology, more and more researchers tried to explore the integration of CNN and SOD. Zhang et al. [26] introduced neural architecture search method to SOD task and proposed the search multi-scale fusion network, which had achieved the function of automatic screening. In [45], a more effective multi-field channel attention module was proposed for deep features to filter channel information more effectively. Wei et al. [38] perceived salient objects in the image through feature fusion, feedback, and multi-level supervision. Pang et al. [28] noticed the sample imbalance in SOD and proposed a consistency-enhanced loss but did not handle the boundary well. Liu et al. [19] introduced Transformer [35] to SOD task. Xu et al. [41] designed a novel knowledge review network to avoid dilution of important information and effectively acquire significant information. Zhao et al. [50] proposed a novel framework to treat semantic context, spatial detail, and boundary information separately.

2.2 Lightweight Salient Object Detection

To achieve a trade-off between practicality and performance, some lightweight SOD models were proposed [30,22,23]. Qin et al. [30] designed a novel and simple network architecture, U^{2}Net, in which a residual u-block was designed to extract intra-stage multi-scale features without degrading the feature map resolution. Liu et al. [22] proposed a lightweight model HVPNet for SOD task, which could imitate the primate visual hierarchies for better multi-scale learning. Liu et al. [23] proposed a lightweight encoder-decoder architecture SAMNet for SOD, and the proposed stereoscopically attentive multi-scale module adopted a stereoscopic attention mechanism for effective and efficient multi-scale learning. Nevertheless, lightweight architectures limit the performance of these models, and a practical and straightforward approach is needed to improve the light-weighted model's weak generalization capability.

2.3 Knowledge Distillation

Large deep neural networks have achieved remarkable success due to their brilliant performance, but the use and deployment of large models on mobile devices is a challenge. To solve this problem, Bucilua et al. [3] first proposed model compression, which transforms the information in a large model or model ensemble into training a small model without causing a significant drop in accuracy. Later, this practice of learning small models from large models was formally generalized as knowledge distillation [9] and was extensively studied on different computer vision tasks [31,15,20]. Shen et al. [31] improved the top-1 accuracy of ResNet-50 to 80%+ through ensemble knowledge distillation without any tricks. Li et al. [15] firstly extended knowledge distillation method to object detection. Liu et al. [20] presented two structured knowledge distillation schemes for semantic segmentation. For salient object detection, the application is limited due to the heavy models and slow detection speed. While some lightweight models have been proposed, the loss of accuracy is intolerable. Inspired by the works mentioned above, we attempt to apply knowledge distillation technique to SOD task, aiming to improve the performance of the lightweight model by profiting from the stronger generalization ability of the heavyweight model.

3 Methodology

The overall architecture of the proposed framework and knowledge distillation process are illustrated in Fig. 2. During distillation, the pre-trained teacher generates weak labels from nature images, and the student model, SCFNet, uses those weak labels to imitate the teacher's generalization ability. The details of the proposed SCFNet and pre-trained teacher model will be elaborated on in the following.

3.1 Backbone Network

We choose MobileNet V3 Large 0.5 [10] as the backbone of SCFNet and remove the fully connected layer to make it meet the SOD task. Assuming that the size of the input image is $H \times W \times 3$, where H and W are height and width, we can obtain five-layer features $\{f^i | i = 1, 2, 3, 4, 5\}$ with sizes $[\frac{H}{2^i}, \frac{W}{2^i}]$ from the backbone network. Features f^1, f^2, and f^3 with larger size are low-level features, f^4 and f^5 are high-level features. Besides, feature f^1 brings much computational

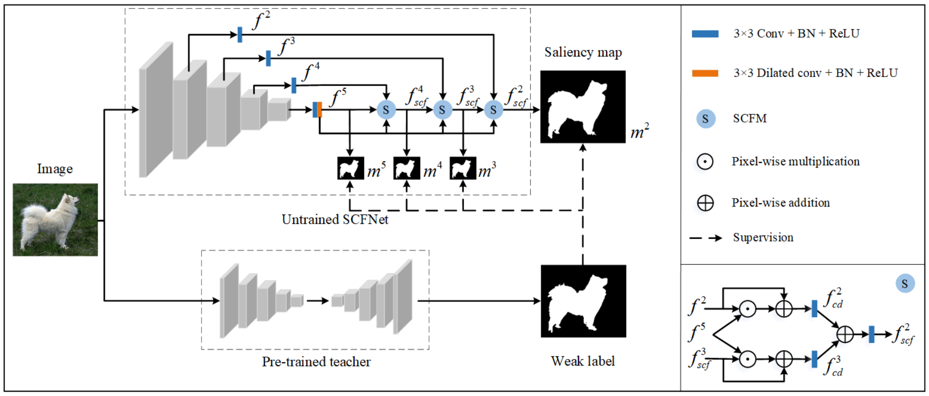

Fig. 2. The overall architecture of the proposed SCFNet and the distillation process. $m^i, i \in \{2, 3, 4, 5\}$ are saliency maps obtained by the processed features f^i. SCFM is the semantics-guided contextual fusion module, and the lower right corner describes the feature processing flow with the last SCFM as an example.

cost and slight performance improvement; hence, we only use the last four-layer features for subsequent processing. Furthermore, all features from the backbone are processed by 3×3 convolution layers to reduce feature dimensions and keep the dimensions consistent. Precisely, we adjust the four-layer features to 16 channels, and a dilated convolution with the dilation ratio of 5 is used for feature f^5 to obtain deeper semantic information.

3.2 Semantics-guided Contextual Fusion Module

The processed features have sufficient size information and the hierarchical information between details and semantics. To make full use of the detailed information and the high-level semantic information, we design a semantics-guided contextual fusion module. Different from the previous works [18,4], and instead of introducing rough boundary in deep features to the fused features, we introduce the location information of objects in deep features into shallow features. Specifically, there are three features sent to the i-th SCFM, namely feature f^i, f^{i+1}_{scf} and f^5, where f^{i+1}_{scf} is the output from the former SCFM, and f^5 is the deepest feature. In the training process, the model should first pay attention to the position information of the object and then refine the rough boundary. Thus, we first perform pixel-wise multiplication operations to transfer the position information of the object in the deep feature f^5 to the shallower features f^i and f^{i+1}_{scf}, and the complex background noises will be suppressed. After 3×3 convolution, a pixel-wise addition operation is used to add the detailed information of the shallow features and obtain the features f^i_{cd} and f^{i+1}_{cd}. Finally, because the contextual features f^i_{cd} and f^{i+1}_{cd} have more prominent object position information and rich object details information, we perform simply pixel-wise addition operation to complete feature fusion, and the fused feature f^i_{scf} are output after

3×3 convolution. For easier understanding, we also describe the above process by taking the last SCFM as an example in the lower right corner of Fig. 2.

3.3 Supervision

The commonly used binary cross-entropy (BCE) function in SOD task calculates the loss of each pixel independently and does not deal well with the sample imbalance issue. The ratio of positive and negative samples in the training set is about 2:5, which leads to poor performance of the model supervised by BCE loss. Some works [25,27,34] applied DICE loss to solve the sample imbalance between the foreground and background areas in image segmentation. DICE loss imposes global constraints on the results such that the prediction maps are related to the sum of the ground truth. Thus, we introduce DICE loss [27] to overcome the sample imbalance caused by scale variation, and the formula is as follow:

$$L_{dice}(P,G) = 1 - \frac{1 + \sum_{i=1,j=1}^{H,W} 2 \times G_{ij} \times P_{ij}}{1 + \sum_{i=1,j=1}^{H,W} G_{ij} + P_{ij}} \tag{1}$$

Although DICE loss can better optimize the model, the model still lacks sufficient constraints on the boundary area, resulting in blurred boundaries in saliency maps. To solve this problem, we construct a boundary DICE (BD) loss to explicitly learn the boundary of salient object, which can be formulated as

$$L_{bd}(P^b,G^b) = 1 - \frac{1 + \sum_{i=1,j=1}^{H,W} 2 \times G^b_{ij} \times P^b_{ij}}{1 + \sum_{i=1,j=1}^{H,W} G^b_{ij} + P^b_{ij}} \tag{2}$$

where

$$P^b_{ij} = \max(P^{b,thin}_{A_{ij}}) \tag{3}$$

$$G^b_{ij} = \max(G^{b,thin}_{A_{ij}}) \tag{4}$$

we first apply a dilation operation and an erosion operation to obtain the boundary maps $G^{b,thin}$ and $P^{b,thin}$ with thin boundaries of the ground truth G and the prediction P according to [2]. Then, we use max-pooling operation to enlarge the coverage area of the boundary. $\max(\cdot)$ means the max-pooling operation, and A_{ij} represents the pooling area that surrounds the pixel (i,j). The larger the area, the thicker the boundary. For thinner boundary maps, the large boundary loss fluctuation in the early training stage is hard to effectively guide the model, and the thicker boundary is difficult to outline the object accurately. Therefore, we propose a progressive boundary supervision method (PBSM). Specifically, the total training epochs are 69. We set the pooling area as 13×13 to get a thicker boundary in the first 10 epochs of training to assist the network to localize the salient object better, and decrease the side length of the pooling area by 2 every 10 epochs. At the last 9 epochs, we use the thinnest boundary to outline the salient object delicately. We combine DICE loss and BD loss and name it structure polishing (SP) loss, which can be summarized as

$$L_{sp} = L_{dice} + \lambda L_{bd} \tag{5}$$

where λ is a hyperparameter that balances the contributions of the two losses. For the sake of simplicity, it is set to 1.

For FS-SCFNet, we use SP loss to guide the model, and for KD-SCFNet, because the weak labels tend not to have meticulous boundaries, we choose DICE as the loss function. The multi-level supervision strategy [38] is used to guide the training model. Specifically, the total outputs of the training model are the saliency maps $m^i (i = 2, 3, 4, 5)$ with the value range [0-1] obtained by features $f^i_{scf}, (i = 2, 3, 4)$ and f^5 after 3×3 convolution and sigmoid operations. We define the sum of the loss as the total loss. In addition, the saliency map m^2 is the result of the model, providing the dominant loss; other saliency maps provide auxiliary loss. And since the value of the auxiliary loss is larger than the dominant loss, we assign them a smaller weight. The total loss is defined as:

$$L_{FS} = \sum\nolimits_{i=2}^{5} \frac{1}{2^{i-2}} L_{sp}(m^i, G) \tag{6}$$

$$L_{KD} = \sum\nolimits_{i=2}^{5} \frac{1}{2^{i-2}} L_{dice}(m^i, G) \tag{7}$$

where L_{KD} is used for weakly supervised training, and L_{FS} is for fully supervised training.

4 Experiments

4.1 Datasets and Evaluation Metrics

We conduct experiments on six datasets, including benchmark SOD datasets DUTS [37], ECSSD [42], HKU-IS [13], DUT-OMRON [43], PASCAL-S [16] and the KD-SOD80K provided by this paper. DUTS dataset, composed of DUTS-TR dataset and DUTS-TE dataset, is currently the largest salient object detection dataset, consisting of 10,553 images used for training and 5,019 images used for testing. ECSSD contains 1,000 images that are semantically meaningful but structurally complex. HKU-IS has 4,447 images, and most of them have multiple disconnected foreground objects with low color contrast. DUT-OMRON includes 5,168 images with one or more salient objects in each image. The PASCAL-S dataset consists of 850 images with cluttered backgrounds and complex foreground objects. In the process of knowledge distillation, we need an additional database of images in natural scenes to generate weak labels to train the student model. Therefore, we choose 80K images from ImageNet [6] dataset under considering the gain and efficiency of distillation, each with at least one salient object. We name this unlabeled dataset KD-SOD80K, and for making a fair comparison, KD-SOD80K does not contain images from SOD datasets mentioned above.

The performances of FS-SCFNet, KD-SCFNet and other state-of-the-art methods are evaluated by seven different evaluation metrics, including parameters (Params), floating point operations (FLOPs), frames per second (FPS), multiply-accumulate operations (MACCs), mean absolute error (M) [29], F-measure (F_β) [1] and E-measure (E_m) [7]. For a fair comparison, all metrics are calculated on the same computer with one Intel i7-11700 CPU and one RTX 3080TI GPU. The input image is resized to the size reported in the corresponding paper when calculating Params, FPS, FLOPs, and MACCs, and other evaluation metrics are calculated by the saliency maps provided by the authors.

4.2 Implementation Details

Consistent with most existing methods [39,28,45], we use ResNet-50 [8] as the backbone network of the teacher model, and the four-layer features from ResNet-50 are adjusted to 64 channels. The rest decoding part is the same as SCFNet. We name this teacher model as SCFNet$_{R50}$.

We use DUTS-TR [37] as the training dataset for the teacher and fully supervised SCFNet. For knowledge distillation based SCFNet, the predicted saliency maps of the teacher model on KD-SOD80K dataset are regarded as the weakly supervised labels. We adopt random flip, random crop, and multi-scale strategy as data enhancement techniques to avoid overfitting problems in the training phase. To ensure model convergence, our network is trained for 69 epochs with a mini-batch of 128. The backbone network is initialized with the corresponding model pre-trained on ImageNet [6] dataset, and the rest is initialized by default. Throughout the whole training process, we use the momentum stochastic gradient descent optimizer with a weight decay of 5e-4 and a momentum of 0.9 to update the weight of the model. The learning rate warm-up and cosine annealing decay method are used. In detail, the learning rate gradually increases from 6.4e-4 to 6.4e-2 at the first five epochs, and then decreases to 6.4e-4 until the end of training. The images are simply resized to 352×352 to predict saliency maps without any post-processing during the testing process.

4.3 Comparison with State-of-the-art Methods

We compare our methods with 3 existing state-of-the-art lightweight methods that we can find, namely U^{2}Net [30], HVPNet [22] and SAMNet [23]. Furthermore, we also compare them with 12 state-of-the-art heavyweight ResNet-50-based methods, namely CPD [40], PoolNet [18], ITSD [51], GateNet [49], DFI [17], MINet [28], GCPANet [4], F^{3}Net [38], PurNet [14] MSFNet [26], SGLKRN [41] and CTDNet [50]. The results of ResNet-50-based SCFNet$_{R50}$ are also reported. The saliency maps are provided by the authors.

Quantitative Comparison. Table 1 shows the quantitative comparison on five popular datasets in terms of Params, FLOPs, M, F_β and E_β. It can be seen that the proposed KD-SFNet outperforms all the lightweight methods across five datasets. Compared with FS-SCFNet, the mean absolute error of KD-SCFNet on the five datasets is reduced by 10.9%, which proves the effectiveness of the proposed distillation method. Furthermore, KD-SCFNet has significant performance advantages compared to other lightweight methods. The results of KD-SCFNet in the mean absolute error on five datasets are reduced by 32.3%, 32.1%, and 24.6%, respectively, compared with SAMNet, HVPNet and U^{2}Net. As a fully supervised lightweight model, although FS-SCFNet is not excellent as KD-SCFNet, it still posses an apparent advantage on accuracy and efficiency over other existing lightweight models. Compared to the heavyweight methods, the lightweight SCFNets contain merely 0.8M parameters and 0.25G FLOPs but still achieves eximious effects. The results by KD-SCFNet on five datasets are even better than those produced by some heavy models, such as PurNet [14], GCPANet [4] and DFI [17]. We also list the comparison between the heavyweight ResNet-50-based models, our SCFNet$_{R50}$ achieves the best results in terms of efficiency and effectiveness. In conclusion, KD-SCFNet and FS-SCFNet provide more accurate and efficient solutions for saliency detection in a weakly supervised and fully supervised way. As a heavyweight one, the proposed SCFNet$_{R50}$ supplies a better choice towards precise SOD.

Table 1. Performance comparison between the proposed methods and the state-of-the-art methods. ↓ (↑) means that the higher (lower) is better. The top two results of lightweight methods and heavyweight methods are shown in red and blue, respectively.

Method	Params	FLOPs	DUTS-TE			ECSSD			HKU-IS			DUT-OMRON			PASCAL-S		
	(M)	(G)	$M\downarrow$	$F_\beta\uparrow$	$E_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\beta\uparrow$
Lightweight methods																	
U^{2}Net$_{\mathrm{PR20}}$	1.13	24.13	0.054	0.818	0.861	0.041	0.916	0.919	0.037	0.898	0.939	0.060	0.784	0.855	0.089	0.830	0.832
HVPNet$_{\mathrm{TCYB21}}$	1.24	1.24	0.058	0.799	0.863	0.052	0.899	0.924	0.044	0.881	0.937	0.065	0.768	0.847	0.093	0.822	0.848
SAMNet$_{\mathrm{TIP21}}$	1.33	0.60	0.058	0.795	0.864	0.050	0.898	0.930	0.045	0.880	0.937	0.065	0.765	0.847	0.095	0.819	0.847
FS-SCFNet	0.80	0.25	0.041	0.849	0.920	0.039	0.923	0.946	0.031	0.917	0.955	0.059	0.785	0.872	0.068	0.860	0.902
KD-SCFNet	0.80	0.25	0.035	0.868	0.932	0.035	0.929	0.952	0.029	0.922	0.959	0.052	0.800	0.882	0.061	0.871	0.912
Heavyweight methods																	
CPD$_{\mathrm{CVPR19}}$	47.85	17.75	0.043	0.843	0.904	0.037	0.924	0.949	0.034	0.904	0.950	0.056	0.780	0.873	0.074	0.852	0.885
PoolNet$_{\mathrm{CVPR19}}$	68.26	49.59	0.040	0.851	0.904	0.039	0.923	0.945	0.032	0.908	0.954	0.056	0.786	0.869	0.076	0.856	0.876
ITSD$_{\mathrm{CVPR20}}$	26.18	23.82	0.041	0.855	0.898	0.034	0.928	0.932	0.031	0.911	0.953	0.061	0.793	0.867	0.066	0.858	0.866
GateNet$_{\mathrm{ECCV20}}$	128.63	136.22	0.040	0.855	0.903	0.040	0.925	0.943	0.033	0.909	0.953	0.055	0.791	0.868	0.071	0.858	0.882
DFI$_{\mathrm{TIP20}}$	29.63	26.96	0.039	0.857	0.908	0.035	0.929	0.949	0.031	0.911	0.957	0.055	0.793	0.870	0.066	0.867	0.893
MINet$_{\mathrm{CVPR20}}$	162.38	105.34	0.037	0.860	0.917	0.033	0.931	0.953	0.029	0.917	0.960	0.056	0.789	0.873	0.066	0.860	0.897
GCPANet$_{\mathrm{AAAI20}}$	67.06	65.72	0.038	0.858	0.913	0.035	0.925	0.951	0.030	0.915	0.957	0.056	0.788	0.868	0.063	0.861	0.900
F^{3}Net$_{\mathrm{AAAI20}}$	25.54	16.43	0.035	0.867	0.918	0.033	0.933	0.946	0.028	0.918	0.958	0.053	0.794	0.876	0.064	0.865	0.892
PurNet$_{\mathrm{TIP21}}$	35.53	48.44	0.039	0.854	0.915	0.035	0.928	0.953	0.030	0.914	0.956	0.051	0.794	0.876	0.070	0.857	0.890
MSFNet$_{\mathrm{ACM21}}$	32.50	19.25	0.034	0.869	0.931	0.033	0.934	0.954	0.027	0.918	0.959	0.050	0.791	0.876	0.063	0.867	0.902
SGLKRN$_{\mathrm{AAAI21}}$	68.69	102.18	0.034	0.878	0.926	0.036	0.932	0.942	0.028	0.923	0.959	0.049	0.810	0.889	0.070	0.871	0.880
CTDNet$_{\mathrm{ACM21}}$	24.64	12.35	0.034	0.876	0.928	0.032	0.936	0.949	0.027	0.926	0.961	0.052	0.807	0.884	0.064	0.870	0.898
SCFNet$_{\mathbf{R50}}$	26.15	13.15	0.031	0.887	0.936	0.031	0.940	0.955	0.026	0.928	0.959	0.047	0.811	0.880	0.060	0.872	0.903

Furthermore, we also illustrate the comparison in Fig. 3, where the trade-offs between accuracy and efficiency of the existing methods are more clearly shown. In the sub-figures of F-measure vs. Params, F-measure vs. FLOPs, F-measure vs. FPS and F-measure vs. MACCs. FS-SCFNet and KD-SCFNet always lie at the top, and achieve excellent performance with 0.8M and 174 FPS real-time detection speed. This implies that our methods make a good trade-off between accuracy and efficiency.

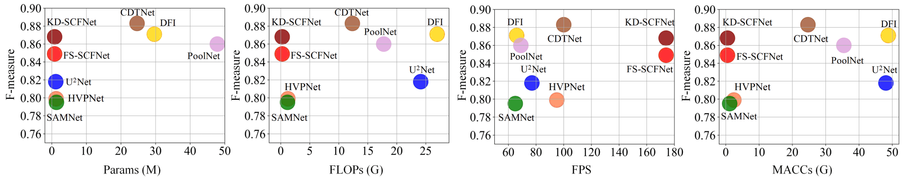

Fig. 3. Illustration of the trade-offs between the accuracy and efficiency of existing models and proposed models. For the sake of clarity, we select all lightweight and some representative heavyweight models for comparison. F-measure is the average of the five testing datasets.

Visual Comparison. We select some representative salient object detection scenes for visual comparison in Fig. 4. These scenes reflect various scenarios, including simple situation (1^{st} row), small objects in low contrast scenes (2^{st} and 3^{st} rows), large object in low contrast scene (4^{st} row), multiple objects with complex structures (5^{st} and 6^{st} rows). It can be seen that the proposed FS-SCFNet and KD-SCFNet can consistently generate accurate and complete saliency maps with sharp boundaries and coherent details.

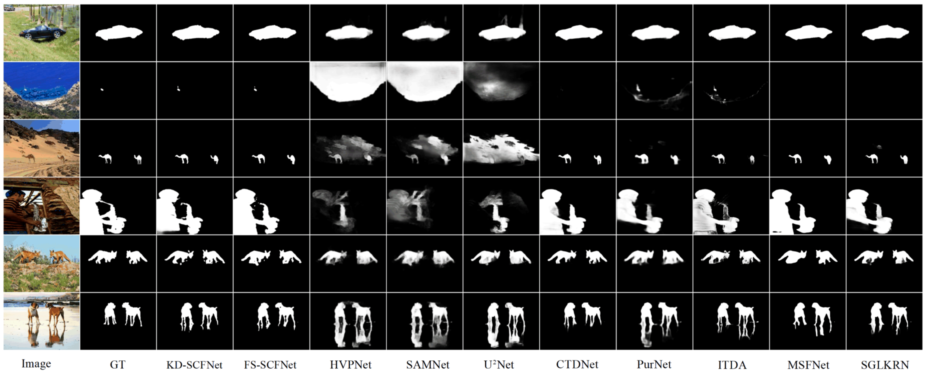

Fig. 4. Visual comparison of our lightweight models and some state-of-the-art models.

4.4 Ablation Study

To demonstrate the effectiveness of the proposed distillation method and modules, we design the following experiments, and all experiments are based on DUTS-TE and ECSSD datasets in terms of the mean absolute error (M), F-measure (F_β) and E-measure (E_β) evaluation metrics.

Effectiveness of Proposed Distillation Method. To verify the effectiveness of the proposed distillation method, we distill different students using SCFNet$_{R50}$ as the teacher and KD-SOD80K as the knowledge distillation dataset. The results are shown in Table 2. It can be seen that the student has a visible improvement in terms of performance, whether it is a lightweight or heavyweight model. In particular, the distilled GCPANet [4] performs better than the teacher model and achieves unprecedented results, which means that the teacher model would not limit the performance of the student model. Fig. 5 shows the comparison between the distillation results and fully supervised results, where the effect of knowledge distillation can be seen more intuitively.

To further explore the influence of different teachers on the same student, we distill the student SCFNet using different teachers. In detail, we change the backbone network of SCFNet to VGG-16 [33] and Swin Transformer-B [24] to obtain different teachers, and represent them as SCFNet$_{V16}$ and SCFNet$_{STB}$, respectively. In addition, we also use an existing method CTDNet [50] as the teacher model, which has an absolutely different architecture from SCFNet. As shown in Fig. 6, KD-SCFNet has an obvious promotion with the improvement of the teacher models, which means that better teacher can more accurately deliver the knowledge to student. Besides, even though the CTDNet with an absolutely different structure is used as the teacher, KD-SCFNet still achieves excellent results, which proves the robustness of the proposed distillation method.

Table 2. Distillation results for different models. We show the results of knowledge distillation based models, like KD-U^{2}Net, KD-SAMNet, KD-GCPANet and the original results. The best results are shown in red.

Model	DUTS-TE			ECSSD		
	$M \downarrow$	$F_\beta \uparrow$	$E_\beta \uparrow$	$M \downarrow$	$F_\beta \uparrow$	$E_\beta \uparrow$
U^{2}Net	0.054	0.818	0.861	0.041	0.916	0.919
KD-U^{2}Net	0.049	0.841	0.904	0.037	0.927	0.950
SAMNet	0.058	0.795	0.864	0.050	0.898	0.930
KD-SAMNet	0.053	0.811	0.895	0.042	0.909	0.943
GCPANet	0.038	0.858	0.913	0.035	0.925	0.951
KD-GCPANet	0.034	0.878	0.934	0.029	0.938	0.959

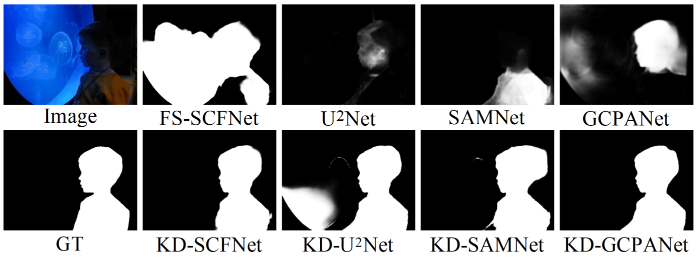

Fig. 5. Visual comparison of the distilled models and the fully supervised models.

Effectiveness of Proposed Modules. To prove the effectiveness of the proposed SCFM, SP loss, and PBSM, the following experiments are designed under the fully supervised training. First, we construct a BASE model consisting of a MobileNet V3 encoder and a basic decoder to compare with the proposed SCFM. In this decoder, the multi-level features from the encoder are concatenated channel-wise and compressed into a saliency map by a 3×3 convolution layer. Then, we use different loss functions to supervise those models. Specifically, we use BCE and DICE, respectively, to compare with the proposed SP loss. Finally, we fix the pooling areas of SP loss to be 13×13, 5×5, and 1×1 to obtain thicker, thick, and thin boundaries, to verify the effectiveness of PBSM. The results are shown in Table 3. It is clearly seen that both SCFM and SP loss can effectively improve the model's performance. Additionally, compared with using boundaries with fixed thicknesses, PBSM acts as an auxiliary tool to guide the proposed network to learn better by alleviating the boundary bias between the prediction map and the ground truth.

Table 3. Results of ablation experiments on SCFM and SP loss. We fix the pooling area to 13×13, 5×5, 1×1, respectively, to demonstrate the effectiveness of the proposed PBSM. SP$_{pbsm}$ means SP loss with progressive boundary supervision method. The best results are shown in red.

Model & Loss	DUTS-TE			ECSSD		
	$M\downarrow$	$F_\beta\uparrow$	$E_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\beta\uparrow$
BASE & BCE	0.065	0.760	0.831	0.066	0.862	0.901
FS-SCFNet & BCE	0.055	0.800	0.856	0.055	0.889	0.917
FS-SCFNet & DICE	0.045	0.832	0.914	0.043	0.911	0.941
FS-SCFNet & SP$_{13\times13}$	0.043	0.839	0.916	0.042	0.918	0.939
FS-SCFNet & SP$_{5\times5}$	0.041	0.848	0.918	0.041	0.921	0.943
FS-SCFNet & SP$_{1\times1}$	0.042	0.843	0.917	0.041	0.920	0.942
FS-SCFNet & SP$_{pbsm}$	0.041	0.849	0.920	0.039	0.923	0.946

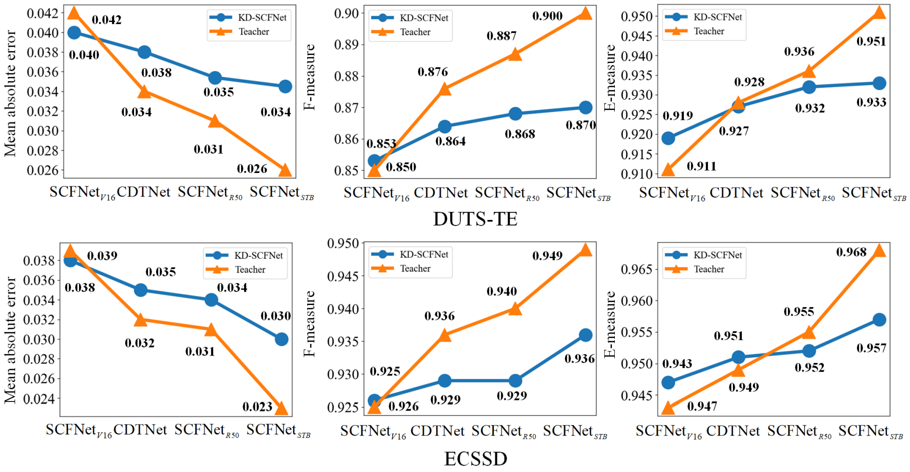

Fig. 6. Distillation results of KD-SCFNet under different teacher models. The teachers are SCFNet$_{V16}$, CTDNet, SCFNet$_{R50}$ and SCFNet$_{STB}$, respectively.

5 Conclusion

In this work, we strive to face the challenge of constructing a SOD model that achieves good accuracy and efficiency. We first design a semantics-guided contextual fusion module (SCFM) that can effectively utilize contextual detail information and global semantic information. Embarking on SCFM, we propose a novel lightweight semantics-guided contextual fusion network (SCFNet), which endows with more accurate and efficient than existing lightweight SOD models. Moreover, we introduce knowledge distillation to SOD task to heighten the generalization ability of lightweight models in a weakly supervised manner. The distilled KD-SCFNet achieves comparable accuracy to the state-of-the-art heavyweight models while maintaining much faster speed, much fewer parameters, and FLOPs. Furthermore, we design a novel structure polishing loss and a progressive boundary supervision method to mitigate the adverse effects of thinner or thicker boundaries. Numerous experiments have demonstrated the effectiveness of the proposed distillation method and framework.

Acknowledgements

This work is supported in part by grants from the National Natural Science Foundation of China (No. 61806126, 61903256, 61976140, 61973307, 62062040), the Natural Science Foundation of Shanghai (19ZR1455300, 21ZR1462600), the Shanghai Science and Technology Innovation Action Plan (No. 22S31903900), the Outstanding Youth Project of Jiangxi Natural Science Foundation (No. 20212ACB212003) and the Jiangxi Province Key Subject Academic and Technical Leader Funding Project (No. 20212BCJ23017). Many thanks for the GPUs provided by Baidu PaddlePaddle AI Studio.

References

1. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: 2009 IEEE conference on computer vision and pattern recognition (CVPR). pp. 1597–1604. IEEE (2009)
2. Bokhovkin, A., Burnaev, E.: Boundary loss for remote sensing imagery semantic segmentation. In: International Symposium on Neural Networks. pp. 388–401. Springer (2019)
3. Bucilua, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 535–541 (2006)
4. Chen, Z., Xu, Q., Cong, R., Huang, Q.: Global context-aware progressive aggregation network for salient object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 10599–10606 (2020)
5. Cheng, M., Mitra, N.J., Huang, X., Torr, P.H., Hu, S.: Global contrast based salient region detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(3), 569–582 (2014)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
7. Fan, D., Gong, C., Cao, Y., Ren, B., Cheng, M., Borji, A.: Enhanced-alignment Measure for Binary Foreground Map Evaluation. In: International Joint Conference on Artificial Intelligence (IJCAI). pp. 698–704 (2018), http://dpfan.net/e-measure/
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
10. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1314–1324 (2019)
11. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on pattern analysis and machine intelligence **20**(11), 1254–1259 (1998)
12. Jiang, L., Xu, M., Wang, X., Sigal, L.: Saliency-guided image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16509–16518 (2021)
13. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5455–5463 (2015)
14. Li, J., Su, J., Xia, C., Ma, M., Tian, Y.: Salient object detection with purificatory mechanism and structural similarity loss. IEEE Transactions on Image Processing **30**, 6855–6868 (2021)

15. Li, Q., Jin, S., Yan, J.: Mimicking very efficient network for object detection. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 6356–6364 (2017)
16. Li, Y., Hou, X., Koch, C., Rehg, J.M., Yuille, A.L.: The secrets of salient object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 280–287 (2014)
17. Liu, J., Hou, Q., Cheng, M.: Dynamic feature integration for simultaneous detection of salient object, edge, and skeleton. IEEE Transactions on Image Processing **29**, 8652–8667 (2020)
18. Liu, J., Hou, Q., Cheng, M., Feng, J., Jiang, J.: A simple pooling-based design for real-time salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3917–3926 (2019)
19. Liu, N., Zhang, N., Wan, K., Han, J., Shao, L.: Visual saliency transformer. arXiv preprint arXiv:2104.12099 (2021)
20. Liu, Y., Shu, C., Wang, J., Shen, C.: Structured knowledge distillation for dense prediction. IEEE transactions on pattern analysis and machine intelligence (2020)
21. Liu, Y., Cheng, M.M., Zhang, X.Y., Nie, G.Y., Wang, M.: DNA: Deeply supervised nonlinear aggregation for salient object detection. IEEE Transactions on Cybernetics (2021)
22. Liu, Y., Gu, Y.C., Zhang, X.Y., Wang, W., Cheng, M.M.: Lightweight salient object detection via hierarchical visual perception learning. IEEE Transactions on Cybernetics (2020)
23. Liu, Y., Zhang, X.Y., Bian, J.W., Zhang, L., Cheng, M.M.: SAMNet: Stereoscopically attentive multi-scale network for lightweight salient object detection. IEEE Transactions on Image Processing **30**, 3804–3814 (2021)
24. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. International Conference on Computer Vision (ICCV) (2021)
25. Mac, B., Moody, A.R., Khademi, A.: Siamese content loss networks for highly imbalanced medical image segmentation. In: Medical Imaging with Deep Learning. pp. 503–514. PMLR (2020)
26. Miao, Z., Liu, T., Piao, Y., Yao, S., Lu, H.: Auto-msfnet: Search multi-scale fusion network for salient object detection. In: ACM Multimedia Conference 2021 (2021)
27. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. IEEE (2016)
28. Pang, Y., Zhao, X., Zhang, L., Lu, H.: Multi-scale interactive network for salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9413–9422 (2020)
29. Perazzi, F., Krähenbühl, P., Pritch, Y., Hornung, A.: Saliency filters: Contrast based filtering for salient region detection. In: 2012 IEEE conference on computer vision and pattern recognition (CVPR). pp. 733–740. IEEE (2012)
30. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O.R., Jagersand, M.: U2-net: Going deeper with nested u-structure for salient object detection. Pattern Recognition **106**, 107404 (2020)
31. Shen, Z., Savvides, M.: Meal v2: Boosting vanilla resnet-50 to 80%+ top-1 accuracy on imagenet without tricks. arXiv preprint arXiv:2009.08453 (2020)
32. Shimoda, W., Yanai, K.: Weakly supervised semantic segmentation using distinct class specific saliency maps. Computer Vision and Image Understanding **191**, 102712 (2020)

33. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
34. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep learning in medical image analysis and multimodal learning for clinical decision support, pp. 240–248. Springer (2017)
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
36. Wang, H., Li, Z., Li, Y., Gupta, B.B., Choi, C.: Visual saliency guided complex image retrieval. Pattern Recognition Letters **130**, 64–72 (2020)
37. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 136–145 (2017)
38. Wei, J., Wang, S., Huang, Q.: F^{3}net: Fusion, feedback and focus for salient object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12321–12328 (2020)
39. Wei, J., Wang, S., Wu, Z., Su, C., Huang, Q., Tian, Q.: Label decoupling framework for salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13025–13034 (2020)
40. Wu, Z., Su, L., Huang, Q.: Cascaded partial decoder for fast and accurate salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3907–3916 (2019)
41. Xu, B., Liang, H., Liang, R., Chen, P.: Locate globally, segment locally: A progressive architecture with knowledge review network for salient object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 3004–3012 (2021)
42. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1155–1162 (2013)
43. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3166–3173 (2013)
44. Yao, Y., Chen, T., Xie, G.S., Zhang, C., Shen, F., Wu, Q., Tang, Z., Zhang, J.: Non-salient region object mining for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2623–2632 (2021)
45. Zhang, J., Shi, Y., Zhang, Q., Cui, L., Chen, Y., Yi, Y.: Attention guided contextual feature fusion network for salient object detection. Image and Vision Computing p. 104337 (2021)
46. Zhang, J., Ehinger, K.A., Ding, J., Yang, J.: A prior-based graph for salient object detection. In: 2014 IEEE international conference on image processing (ICIP). pp. 1175–1178. IEEE (2014)
47. Zhang, P., Zhuo, T., Huang, W., Chen, K., Kankanhalli, M.: Online object tracking based on cnn with spatial-temporal saliency guided sampling. Neurocomputing **257**, 115–127 (2017)
48. Zhang, P., Liu, W., Wang, D., Lei, Y., Wang, H., Lu, H.: Non-rigid object tracking via deep multi-scale spatial-temporal discriminative saliency maps. Pattern Recognition **100**, 107130 (2020)

49. Zhao, X., Pang, Y., Zhang, L., Lu, H., Zhang, L.: Suppress and balance: A simple gated network for salient object detection. In: European Conference on Computer Vision. pp. 35–51. Springer (2020)
50. Zhao, Z., Xia, C., Xie, C., Li, J.: Complementary trilateral decoder for fast and accurate salient object detection. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 4967–4975 (2021)
51. Zhou, H., Xie, X., Lai, J.H., Chen, Z., Yang, L.: Interactive two-stream decoder for accurate and fast saliency detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9141–9150 (2020)