Disentangle Saliency Detection into Cascaded Detail Modeling and Body Filling

YUE SONG*, University of Trento, Italy
HAO TANG*, ETH Zurich, Switzerland
NICU SEBE, University of Trento, Italy
WEI WANG, University of Trento, Italy

Salient object detection has been long studied to identify the most visually attractive objects in images/videos. Recently, a growing amount of approaches have been proposed all of which rely on the contour/edge information to improve detection performance. The edge labels are either put into the loss directly or used as extra supervision. The edge and body can also be learned separately and then fused afterward. Both methods either lead to high prediction errors near the edge or cannot be trained in an end-to-end manner. Another problem is that existing methods may fail to detect objects of various sizes due to the lack of efficient and effective feature fusion mechanisms. In this work, we propose to decompose the saliency detection task into two cascaded sub-tasks, *i.e.*, detail modeling and body filling. Specifically, the detail modeling focuses on capturing the object edges by supervision of explicitly decomposed detail label that consists of the pixels that are nested on the edge and near the edge. Then the body filling learns the body part which will be filled into the detail map to generate more accurate saliency map. To effectively fuse the features and handle objects at different scales, we have also proposed two novel multi-scale detail attention and body attention blocks for precise detail and body modeling. Experimental results show that our method achieves state-of-the-art performances on six public datasets.

CCS Concepts: • **Computing methodologies → Scene understanding**.

Additional Key Words and Phrases: Salient Object Detection, Visual Saliency, Foreground Segmentation

1 INTRODUCTION

Human Visual System (HVS) has the innate ability to capture salient objects from visual scenes rapidly without training [31]. Salient Object Detection (SOD) aims at simulating HVS to detect distinctive regions or objects, where people would like to focus their eyes on [2, 34]. In the past decades, it has attracted much interest from research communities, mainly because it can find objects or regions that can represent a scene efficiently, a useful step in downstream computer vision tasks. Saliency detection models have been evolved from traditional hand-engineering approaches via different saliency cues (*e.g.*, global contrast [6], background prior [42], and spectral analysis [16]) to Fully Convolutional Neural Networks (FCN) [21] based methods.

Despite that FCN-based solutions [15, 20, 25, 27, 29, 36, 37, 43, 47, 50] have made remarkable progress so far, there still exist two main challenges: (i) the pixels near the object edge have a very imbalanced distribution, which makes these pixels harder to predict than the non-edge ones. Existing saliency detection models usually get large

*Both authors contributed equally to this research.

Authors' addresses: Yue Song, yue.song@unitn.it, University of Trento, Italy, 38122; Hao Tang, hao.tang@vision.ee.ethz.ch, ETH Zurich, Switzerland, 8092; Nicu Sebe, nicu.sebe@unitn.it, University of Trento, Italy, 38122; Wei Wang, wei.wang@unitn.it, University of Trento, Italy, 38122.

1551-6857/2022/1-ART1 $15.00
https://doi.org/10.1145/3513134

prediction error when the pixel is close to the object boundary [37]; (ii) most saliency detection methods build models on the encoder-decoder framework and develop different strategies to aggregate multi-scale features for better representation. However, due to the lack of effective fusion mechanisms to integrate multi-scale or multi-level feature, the generated saliency maps may fail to accurately predict objects in different scales. Because of these two issues, existing methods might fail to generate accurate saliency maps with sharp boundaries and coherent details (see Fig. 1).

For the first problem, many methods attempted to introduce boundary information as extra supervision to improve the prediction performances [12, 19, 27, 29, 40, 47]. However, the introduced edge label only indicates the pixel on the edge. Its direct use as supervision can decrease the global saliency prediction error but will degrade the prediction performance near the edge [37]. More recently, Wei *et al.* [37] proposed to explicitly decompose the ground truth saliency label into the body label and the detail label. The decomposed detail label consists of both edges as well as nearby pixels, which makes full use of pixels near the edge and thus has a more balanced pixel distribution. The decoupled body and detail maps are used to train two separate network branches, and a feature fusion branch is needed to combine the two streams to generate the final saliency map. Their proposed architecture involves two iterations *i.e.*, train the detail and body branches until the two branches can output good body/detail maps, and then train the fusion module. It is non-trivial to control the two iterations and train the model in an end-to-end manner.

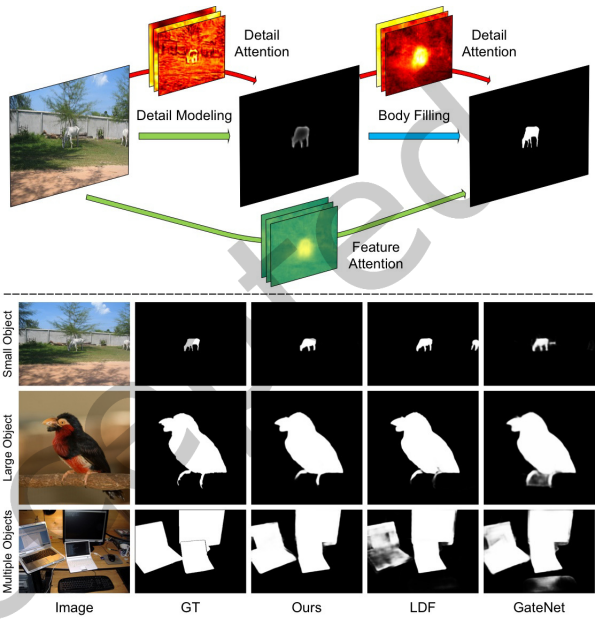

Fig. 1. (**Top**) We disentangle the task of salient object detection into cascaded detail modeling and body filling. The proposed multi-scale attention blocks polish the features passed via short connections and help the network attend to the salient regions. (**Bottom**) Qualitative comparison between our method and two recent state-of-the-art methods LDF [37] and GateNet [50]. Our approach can precisely segment objects of various sizes with subtle details.

For the second problem, some methods tried to pass the features at the corresponding level in the encoder to the decoder via different connection pathways to leverage multi-level context information [4, 12, 22, 27, 34, 38, 39, 43, 46, 51]. Without being processed by proper mechanism, the representation power of details in a single shallow layer may be weakened or disturbed by deeper features with high-level semantic information. To more effectively utilize multi-scale features, some methods proposed to pass multi-layer features to a decoder in a single layer in the fully connected manner or the heuristic style [15, 33, 44]. However, this kind of solution suffers from huge computational burden and fusion difficulties brought by the excessive amount of features and their resolution gap. There lacks such a mechanism that can effectively and efficiently fuse multi-level features without losing representation power at different scales.

To address both aforementioned issues, we propose a novel solution: a cascaded framework that disentangles traditional SOD task into two sub-tasks, where the first sub-network is forced to generate the detail map by supervision of decomposed detail label as proposed in [37], subsequently the second sub-network takes in the detail map, detail feature, and the image to generate the body map and fuse into the final saliency map. The proposed framework explicitly divides the original task into two cascaded sub-tasks each of which has its own

specific target. This reduces the difficulty in directly predicting the whole saliency map. Besides the framework, we also propose two novel multi-scale attention blocks that aim at fusing features at different levels and detecting objects of various sizes. The proposed blocks can enrich the fused feature with multi-scale attention, which can effectively fuse two or three multi-level features and relax the difficulty in the detection of multi-scale objects. In addition, we suggest a hybrid loss setting that targets the accurate generation of each map and can well complement each other. Our model is trained in an end-to-end fashion and has a reasonable inference speed of 20 FPS on a single GPU. The proposed model is thoroughly validated under four metrics across six public benchmark datasets to demonstrate its superior performances.

In summary, the main contributions of the paper are as follows:

- We propose a novel cascaded saliency detection framework that first produces detail maps of the object and then generates accurate saliency map by filling the detail map with body map. The proposed framework reduces the difficulty in directly predicting the whole saliency map and can be trained efficiently in an end-to-end manner.
- We propose two novel multi-scale attention blocks that can attentively fuse multiple features at multiple scales for precise detail and body map generation. We also suggest a hybrid loss setting that specifically targets the detail and body maps and complement each other.
- Our proposed model achieves state-of-the-art performances against 10 most recent state-of-the-art methods on six benchmark datasets under four widely used metrics. Extensive ablation studies are also conducted to demonstrate the effectiveness of each proposed module.

2 RELATED WORK

Early saliency detection methods in hand-engineered era mainly rely on various saliency cues, including global or local contrast [6, 11], background prior [42], and spectral analysis [13, 16]. Due to the page limits, the readers are kindly referred to [2] for a detailed review. Here we recap modern approaches in deep learning era. These FCN-based methods can be broadly divided into two families as follows:

Aggregation-based Models. Most modern saliency detection models are based on the encoder-decoder framework to integrate multi-level features and leverage contextual information across different layers [3, 4, 12, 15, 22, 25, 27, 30, 33, 36, 38, 39, 43–46]. The encoder is often used to extract multi-level features from the image, and the decoder is designed to effectively combine the features and predict the saliency map. During the past years, researchers have developed lots of feature fusion mechanisms and feature connection pathways for better representation. Liu *et al.* [20] proposed a hierarchical pixel-wise contextual attention network to learn the local and global context for each pixel. Zhao *et al.* [49] utilized channel attention for high-level representation and spatial attention for low-level feature maps to improve the detection performances. Our proposed two attention blocks share some similarities with [49] but are fundamentally different in three aspects. First, we do not distinguish channel attention or spatial attention for high-level or low-level features. Instead, we keep consistently using the combination of global and local attention for all the feature maps to be fused, regardless of its layer. Secondly, our attention blocks work in multiple scales and the fused representation would choose the information needed at a certain scale. Lastly, we have different architecture design and can take in two or three feature streams.

Edge-guided Models. In recent years, increasingly more approaches incorporate the edge/contour information to assist SOD task and improve the detection performances [12, 19, 27, 29, 37, 40, 44, 47, 49]. Zhao *et al.* [47] used edge label to supervise low-level feature maps to enable the network to have the capacity of modeling edge information. More recently, Wei *et al.* [37] proposed to explicitly decompose the ground truth label into the detail label that consists of pixels on the edge as well as pixels nearby the edge and the body label that concentrates on the pixels far from the edge. The two decoupled labels are used to supervise two branches in the first iteration, and a second iteration is still needed to train the fusion module for combing the results. Although

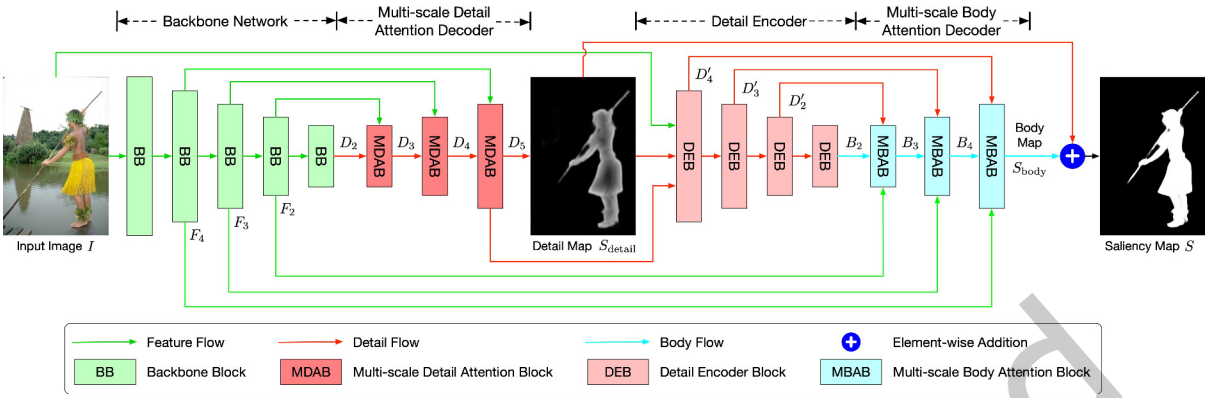

Fig. 2. Overview of the proposed framework. The left sub-network consists of a backbone network and a proposed multi-scale detail attention decoder. The detail decoder takes in the feature flow transmitted from the backbone and outputs detail map. Subsequently, the detail encoder absorbs the input image, generated detail map, and the detail flow from the last MDAB. The fused feature is exploited by detail encoder to produce new detail flow for the body decoder. Then we feed the proposed multi-scale body attention decoder with detail and feature streams, where the contextual information is leveraged to predict the body map. The final saliency map is obtained by summing up the detail and body map.

we also decompose the original label, only the detail label is used to supervise intermediate results in our method. Another key difference is that our model is cascaded and can be trained end-to-end efficiently.

3 METHODOLOGY

We start by introducing how the detail label is decomposed, then describe each part of the model in detail, and end with the loss function.

Detail Label Generation. As stated before, the pixels near the edge are hard to predict and prone to be misclassified. In the saliency detection task, the ground truth label is often binary and all the pixels in the salient regions have a unique value. Inspired by [37], we explicitly decompose the detail label from the ground truth label and use it for our first sub-task detail modeling. More specifically, given the ground truth label G, we use the *distance transformation* to convert the original label into the detail label in which each pixel in the original salient regions is defined by its minimum distance to the object boundary. This *distance transformation* process can be described as:

$$G_{\text{detail}}(p,q) = \begin{cases} |G(p,q) - E(p,q)|, & G(p,q) = 1, \\ 0, & G(p,q) = 0, \end{cases} \tag{1}$$

where G_{detail} represents the detail label, and $E(p,q)$ denotes the salient edge point that has the minimum Euclidean distance to saliency pixel $G(p,q)$. Fig. 3 displays two examples of decomposed detail labels. After decoupling the detail label, it will be used to supervise the left sub-network in Fig. 2 to detect detail points close to the edge.

Feature Extractor. Similar to existing models [36, 37, 47], we use ResNet-50 [14] as our backbone network. We choose ResNet-50 as the backbone because it has the moderate model size and reasonable feature extraction power. Lager models can lead to better performances but will slow down the training and inference. The last fully connected layer and average pooling layer are removed, and we only keep the convolutional blocks. These blocks generate feature maps at five different scales $\{F_i|i{=}1,\ldots,5\}$, where the resolution is down-sampled by two between subsequent blocks. As pointed out in [39], the representation F_5 at the shallowest layer contains too much coarse and redundant information. This increases computational burdens dramatically but brings little

Fig. 3. Some examples of decoupled detail labels. The pixels in decomposed label have larger values closer to the edge and smaller even zero value when far from the edge, which has a more balanced distribution than pure edge label.

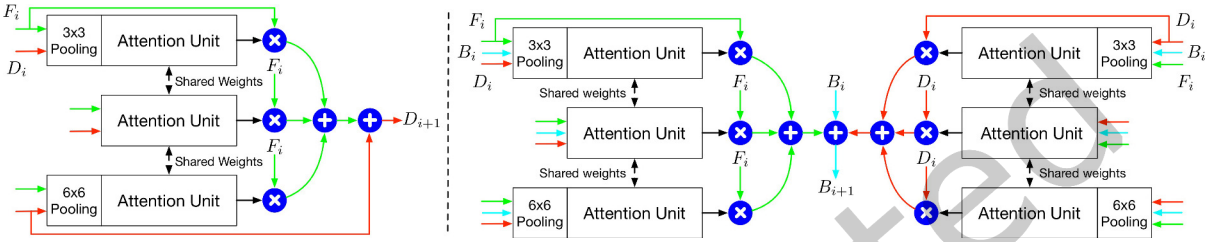

Fig. 4. (**Left**) Workflow of our MDAB. The feature and detail streams are processed by three attention units to leverage contextual information and produce multi-scale attention for feature flow. Then the attentive feature flow is combined with detail flow for the generation of new detail flow. (**Right**) Pipeline of the proposed MBAB. We use three more attention units at different scales to generate attentive feature flow. The new body flow comes out of the summation of attentive detail flow, attentive feature flow, and original body flow.

performance improvement. Hence, we abandon this layer and use only the rest finer features $\{F_i|i{=}1,\ldots,4\}$. Two sets of the features will be passed to the detail decoder and body decoder respectively to assist their tasks.

3.1 Multi-scale Detail Attention Modeling

The detail decoder takes input of the image feature to fulfill the task of detail generation. It consists of three Multi-scale Detail Attention Block (MDAB). As shown in Fig. 4 (*left*), each MDAB absorbs both the detail flow from the block before and the feature flow from the backbone encoder at the corresponding scale, which can be denoted as:

$$D_{i+1} = \mathrm{MDAB}_i(D_i, F_i),\ \ i = 2, 3, 4 \tag{2}$$

where F_i denotes the feature flow at the same level, D_i stands for the current detail flow, and D_{i+1} represents the new detail flow to be passed to the next block. At the last MDAB, we use 3×3 convolution layer and *sigmoid* gate to extract detail map from the final detail flow. This operation can be denoted as, $S_{\mathrm{detail}}{=}\sigma(\mathrm{Conv}(D_5))$, where $\sigma(\cdot)$ denotes the *sigmoid* gate, and $\mathrm{Conv}(\cdot)$ represents the convolution operation. The MDAB is mainly comprised of three attention units, where each one calculates the combination of local and global attention at one scale. The detailed architecture of the proposed attention unit is illustrated in Fig. 5. Inside the attention unit, the detail flow and feature flow first go through one representation sampler to filter out useless noise and keep only the informative features:

$$F_i^{att} = \mathrm{ReLU}(\mathrm{Conv}(D_i) + \mathrm{Conv}(F_i)), \tag{3}$$

where the ReLU gate actively polishes the representation summed by detail flow and feature flow. Then the raw feature attention F_i^{att} will be fed to two branches to obtain the corresponding local and global attention:

$$F_i^{att} = \sigma(\mathrm{Conv}(F_i^{att})) + \sigma(\mathrm{GAP}(\mathrm{Conv}(F_i^{att}))), \tag{4}$$

where GAP denotes the global average pooling module. Assume the feature F_i has the size of $B{\times}C{\times}H{\times}W$, the first term calculates the local spatial-wise attention of size $B{\times}1{\times}H{\times}W$, and the second term computes the global

channel-wise attention of size $B{\times}C{\times}1{\times}1$. As can be seen from the left part of Fig. 4, the feature and detail flow pass three attention units after different pooling layers. We argue that the strategy of pooling with different kernels before the attention unit enables the feature to automatically search for useful information at a certain scale, which will benefit identifying salient objects of various sizes. Notice that the representation samplers of different attention units have shared weights, as we expect the sampler to have the function of filtering out the noise of fused flow, which should be robust against scale variation. In the end, we obtain new attentive detail flow that can identify both the object and its sharp boundaries:

$$D_{i+1} = \mathrm{Conv}(D_i + \sum_{t=1}^{3}(F_{i(t)} \odot F_{i(t)}^{att})) \tag{5}$$

where $\odot$ represents element-wise multiplication, and t denotes the number of scales used in MDAB.

3.2 Multi-scale Body Attention Filling

After the left detail decoder generates the detail map, the detail encoder in the right part of Fig. 2 will be fed with the input image, the detail flow, and the detail map to extract new detail flow $\{D_i^{'}|i{=}1,2,3,4\}$ for the task of body filling. Similar to the backbone network, the detail encoder will pass the detail flows to the body decoder via both normal path and short connections. Afterward, the body decoder takes in the image feature and detail feature to generate the body map. The body decoding task is completed by three subsequent Multi-scale Body Attention Block (MBAB). Each MBAB absorbs three streams, including the feature flow from the backbone network, the detail flow from the detail encoder, and the body flow from the previous block. This procedure can be represented as:

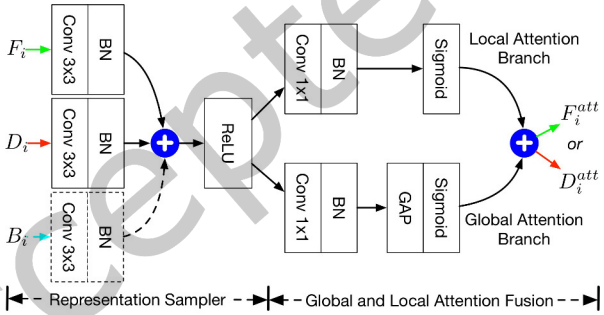

Fig. 5. Architecture of attention unit. The attention unit works as a basic element for our MDAB and MBAB. Depending on the number of input streams, the representation sampler may have different number of convolution blocks. The unit is used to generate the combination of global and local attention for either the feature or detail flow.

$$B_{i+1} = \mathrm{MBAB}(B_i, F_i, D_i), i \in 2,3,4. \tag{6}$$

where B_i is the current body flow, and B_{i+1} is the new body flow passing to the next block. After the three consecutive MBAB, we extract the body map and fill it with the detail map to generate the final saliency prediction $S{=}S_{\mathrm{detail}}{+}S_{\mathrm{body}}$. As we can see in Fig. 4(*right*), our proposed MBAB resembles MDAB, with the main difference in three additional attention units dedicated for the detail flow. The detail and feature flow with multi-scale attention will be fused with body flow to create the new body stream. We can simply denote the MBAB workflow as:

$$B_{i+1} = \mathrm{Conv}(B_i + \sum_{t=1}^{3}(F_{i(t)} \odot F_{i(t)}^{att} + D_{i(t)}^{'} \odot D_{i(t)}^{'att})). \tag{7}$$

3.3 Hybrid Loss Function

As we use both ground truth label and our decomposed detail label to train the network, we design different loss settings for the two tasks. For the detail output, we have the loss function defined as:

$$l_{\mathrm{detail}} = l_{\mathrm{CE}}(S_{\mathrm{detail}}, G_{\mathrm{detail}}) + l_{\mathrm{SSIM}}(S_{\mathrm{detail}}, G_{\mathrm{detail}}), \tag{8}$$

where the first term is the commonly used cross-entropy loss, and the second term is the structural similarity loss which enforces the detail decoder to focus on the edges. Structural similarity index measure (SSIM) [35] was

originally used to calculate the similarity of a pair of images by assessing the structural information. Motivated by [27], we also integrate this loss to let the detail output learn the structural information of the image for keeping the precise edges. This loss is calculated as:

$$l_{\text{SSIM}} = 1 - \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{9}$$

where μ_x, μ_y and σ_x, σ_y are the mean and standard deviation of the image. C_1 and C_2 are small positive constants, and we set them as 0.01^2 and 0.03^2 to avoid dividing zero. As for the body mask prediction, we have the following loss configuration:

$$l_{\text{body}} = l_{\text{CE}}(S, G) + l_{\text{IoU}}(S, G) + l_{\text{F}}(S, G), \tag{10}$$

where the second term is the IoU loss, which can help the body decoder quickly attend to the main body of the object as adopted in [24, 27, 28]. This loss can be computed as:

$$l_{\text{IoU}} = 1 - \frac{\sum_{i=1}^{H}\sum_{j=1}^{W} S(i,j)G(i,j)}{\sum_{i=1}^{H}\sum_{j=1}^{W}(S(i,j) + G(i,j) - S(i,j)G(i,j))}. \tag{11}$$

The third term in Eq. (10) is the so-called F-loss [48]. It is proposed to directly optimize the metric F-measure as defined by: l_{F}=1−$F(S, G)$. Here we expect that adopting this loss can balance the detail and body map to complement the information of each other by pushing their fused mask to achieve a high F-measure score. In total, our model is trained end-to-end using the hybrid loss function:

$$l = \frac{1}{2}(l_{\text{detail}} + l_{\text{body}}). \tag{12}$$

4 EXPERIMENTS

Datasets. Following [36, 37, 39, 40], we conduct extensive experiments on six widely used benchmark datasets to evaluate the effectiveness of the proposed method, *i.e.*, ECSSD [41], PASCAL-S [18], DUT-OMRON [42], HKU-IS [17], THUR15K [5], and DUTS [32].

Specifically, ECSSD [41] contains 1,000 structurally complex natural images. PASCAL-S [18] consists of 850 images with cluttered backgrounds chosen from the validation set of the PASCAL-VOC segmentation dataset [8]. DUT-OMRON [42] has 5,168 images with high content variety. HKU-IS [17] has 4,447 images containing mostly multiple disconnected objects. THUR15K [5] consists of 6,232 diverse and heterogeneous images categorized into several groups. Among these datasets, DUTS [32] is currently the largest saliency detection dataset consisting of two subsets: DUTS-TR contains 10,553 images for training and DUTS-TE has 5,019 images for testing.

Implementation Details. In line with most existing methods [25, 27, 36, 37, 50], we use the DUTS-TR dataset for training and the rest of the datasets as the test set for evaluation. ResNet-50 [14] classifier pre-trained on ImageNet [7] is used as backbone to initialize the model, and the other parameters are randomly initialized. Our network is trained end-to-end for 50 epochs with a mini-batch size of 32 by stochastic gradient descent (SGD). The momentum and weight decay are set to 0.9 and 0.0005, respectively. We set the maximum learning rate to 0.005 for the ResNet-50 backbone and 0.05 for the other parts. Warm-up and linear decay strategies are also used.

During training, we use random horizontal flip, random crop, and multi-scale input images for data augmentation. The images are resized to the resolution of 352×352 during testing and fed into the network to generate the saliency prediction without any post-processing step. Resizing with bilinear interpolation is consistently used throughout all the experiments. The proposed model achieves the inference time of 20 FPS on a single Quadro RTX 6000 GPU.

Evaluation Metrics. We use four widely used metrics to evaluate the proposed method, *i.e.*, Mean Absolute Error (*MAE*) [26], mean F-measure (m F_β) [1], weighted F-measure (F_β^ω) [23], and precision-recall curve. pecifically,

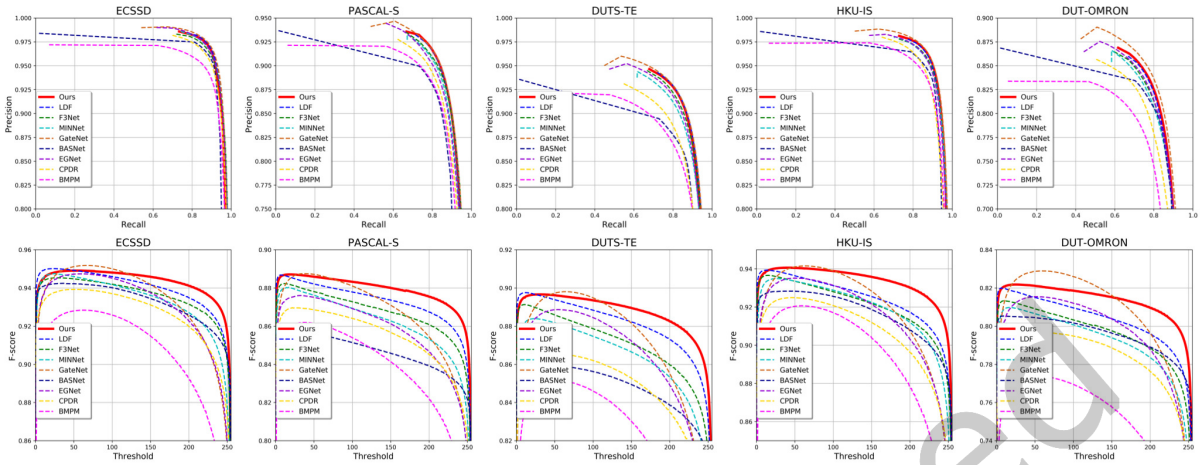

Fig. 6. The precision-recall curve (**first row**) and F-measure versus different thresholds (**second row**) of all the methods.

MAE [26] calculates the absolute per-pixel difference between the saliency prediction and its ground truth:

$$MAE = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} |S(i,j) - G(i,j)|, \tag{13}$$

where W and H is the width and height of the mask, S denotes the predicted saliency map, and G represents the ground truth. As the most fundamental and direct measure, *MAE* has been widely applied to evaluate the quality of saliency map [25, 27, 36, 37, 50]. The generated saliency map S is first converted into a binary map using a threshold and is compared with ground truth G to compute the *precision* and *recall* score:

$$precision = \frac{|S \cap G|}{S}, \ recall = \frac{|S \cap G|}{G} \tag{14}$$

The precision-recall curve is plotted by varying the binarized thresholds from 0 to 255 to obtain a sequence of precision-recalled pairs. The larger the area under the PR curve, the better the performance of the model. F-measure F_β and its weighted variant F_β^ω are used to jointly assess the saliency prediction by taking both *precision* and *recall* into consideration. The basic F-measure can be formulated as:

$$F_\beta = \frac{(1+\beta^2) precision \times recall}{\beta^2 precision + recall}, \tag{15}$$

where β is the relative weight to control the importance of *precision* and *recall*. β is usually set to 0.3 to give a larger weight to *precision* as suggested in [1]. Mean F-measure (m F_β) is computed by taking the mean value of F-measure from the PR curve. Weighted F-measure [23] is an intuitive generalization of F-measure for non-binary maps, which is defined as:

$$F_\beta^\omega = \frac{(1+\beta^2) precision^\omega \times recall^\omega}{\beta^2 precision^\omega + recall^\omega} \tag{16}$$

The basic quantities *precision* and *recall* are extended to non-binary values and assigned different weights to different errors according to location and neighborhood information. Except for these four measures, Max F-measure that selects the maximum value of F-measure from the PR curve, E-measure [10], and S-measure [9] are also used in the literature. S-measure is proposed to compute the region-aware and object-aware similarities, and E-measure is designed to combine local pixel values with image-level mean values for joint assessment.

Table 1. Quantitative results compared with state-of-the-art methods on six datasets. '–' means the results can not be obtained. For all metrics except for *MAE*, higher is better. The best three results are highlighted in **red**, blue, and green respectively.

$\text{Method}_{\text{year}}$	ECSSD (#1,000)			DUTS-TE (#5,019)			DUT-OMRON (#5,168)			PASCAL-S (#850)			HKU-IS (#4,447)			THUR15K (#6,232)		
	MAE	m F_β	F_β^ω	*MAE*	m F_β	F_β^ω	*MAE*	m F_β	F_β^ω	*MAE*	m F_β	F_β^ω	*MAE*	m F_β	F_β^ω	*MAE*	m F_β	F_β^ω
BMPM_{2018} [43]	0.045	0.868	0.871	0.049	0.745	0.761	0.064	0.692	0.681	0.076	0.769	0.782	0.039	0.871	0.859	0.079	0.704	–
CPD-R_{2019} [39]	0.037	0.917	0.898	0.043	0.805	0.795	0.056	0.747	0.719	0.074	0.829	0.800	0.034	0.891	0.875	0.068	0.738	0.730
EGNet-R_{2019} [47]	0.037	0.920	0.903	0.039	0.815	0.816	0.053	0.756	0.738	0.075	0.831	0.807	0.031	0.901	0.887	0.067	0.739	0.733
BANet_{2019} [29]	0.035	0.923	0.908	0.040	0.815	0.811	0.059	0.746	0.736	0.070	0.838	0.817	0.032	0.899	0.887	0.068	0.741	–
BASNet_{2019} [27]	0.037	0.880	0.904	0.048	0.791	0.803	0.056	0.756	0.751	0.079	0.777	0.797	0.032	0.895	0.889	0.073	0.733	0.721
SCRN_{2019} [40]	0.037	0.918	0.899	0.040	0.809	0.803	0.056	0.746	0.720	0.065	0.839	0.816	0.033	0.897	0.878	0.066	0.741	0.734
F3Net_{2020} [36]	**0.033**	0.925	0.912	0.035	0.791	0.835	0.053	0.766	0.747	0.064	0.844	0.823	0.028	0.910	0.900	0.065	0.756	0.744
GateNet_{2020} [50]	0.035	0.917	0.906	0.035	0.816	0.828	0.051	0.761	0.749	0.065	0.827	0.821	0.029	0.903	0.893	–	–	–
MINet_{2020} [25]	**0.033**	0.924	0.911	0.037	0.828	0.825	0.055	0.756	0.738	0.064	0.842	0.821	0.028	0.908	0.899	–	–	–
LDF_{2020} [37]	0.034	0.930	**0.915**	0.034	0.855	0.845	0.051	0.773	0.752	**0.062**	0.853	0.828	**0.027**	0.914	0.904	0.064	0.763	0.752
Ours	0.035	**0.931**	0.911	**0.033**	**0.863**	**0.847**	**0.048**	**0.785**	**0.758**	**0.062**	**0.855**	**0.829**	**0.027**	**0.924**	**0.907**	**0.062**	**0.769**	**0.755**

4.1 State-of-the-Art Comparisons

Quantitative Evaluation. We demonstrate the efficacy of our model by comparing with other 10 most recent state-of-the-art models, including BMPM [43], CPD [39], EGNet [47], BANet [29], BASNet [27], SCRN [40], F3Net [36], GateNet [50], MinNet [25], and LDF [37]. To assure comparison fairness, the saliency maps are either provided by the authors or generated using officially released pre-trained models. Table 1 displays the performances of aforementioned methods on six datasets. Our method consistently outperforms other models and achieves the best performances across six datasets, refreshing the leaderboard and setting the new baseline. In particular, we have significantly improved the best F-score (F_β) over all datasets, with 1.5% increase on DUT-OMRON, 1.1% on HKU-IS, and 0.9% on DUTS-TE. It is also worth mentioning that our method surpasses others by larger margin on large datasets, while the difference on small-scale datasets (<1,000 images) is less obvious. Due to the limited number of images, small datasets may not well reflect the actual performance of a model.

Fig. 6 shows the precision-recall curve (1_{st} row) and the F-measure curve (2_{nd} row) of all the methods. Our PR curve consistently lies above other methods and achieves best performances on ECSSD, PASCAL-S, DUTS-TE, and HKU-IS, and has very competitive results on DUT-OMRON. Moreover, our PR curve is significantly shorter than other methods and has larger recall value ranges, which indicates that our method has less *false negative* predictions in the saliency maps. Across all datasets, our F-measure curve has the flattest slope and largest area under the curve, demonstrating that our generated saliency maps present good quality against varying thresholds.

Notice that on ECSSD and DUT-OMRON the GateNet [50] seem to have the highest F-score at certain thresholds but the score drops very quickly when the threshold is increased. This implies that their saliency maps have many non-binary predictions (*i.e.,* numerical values on (0,1)). Only when an appropriate threshold is carefully chosen, their method could have reasonable performances.

Qualitative Evaluation. Some representative visual examples are shown in Fig. 7. We select images from some challenging scenarios, including low color contrast (1_{st} row), high inter-object contrast (2_{nd} row), low contrast near object boundary (3_{rd} row), multiple objects with low background contrast (4_{th} row), partly occluded object (5_{th} row), object in cluttered backgrounds (6_{th} row), small object near image border (7_{th} row), and object with irregular and complex edges (last row). It can be seen that our method well suppresses background noise and accurately segments the salient objects of various sizes with coherent details.

4.2 Ablation Study

Baseline Models. To investigate the effect of each proposed module, we conduct ablation studies on several baselines to validate the effectiveness of each proposed component: 1) B1 uses ResNet backbone and a decoder network for direct saliency prediction; 2) B2 first generates intermediate body map and then produces detail

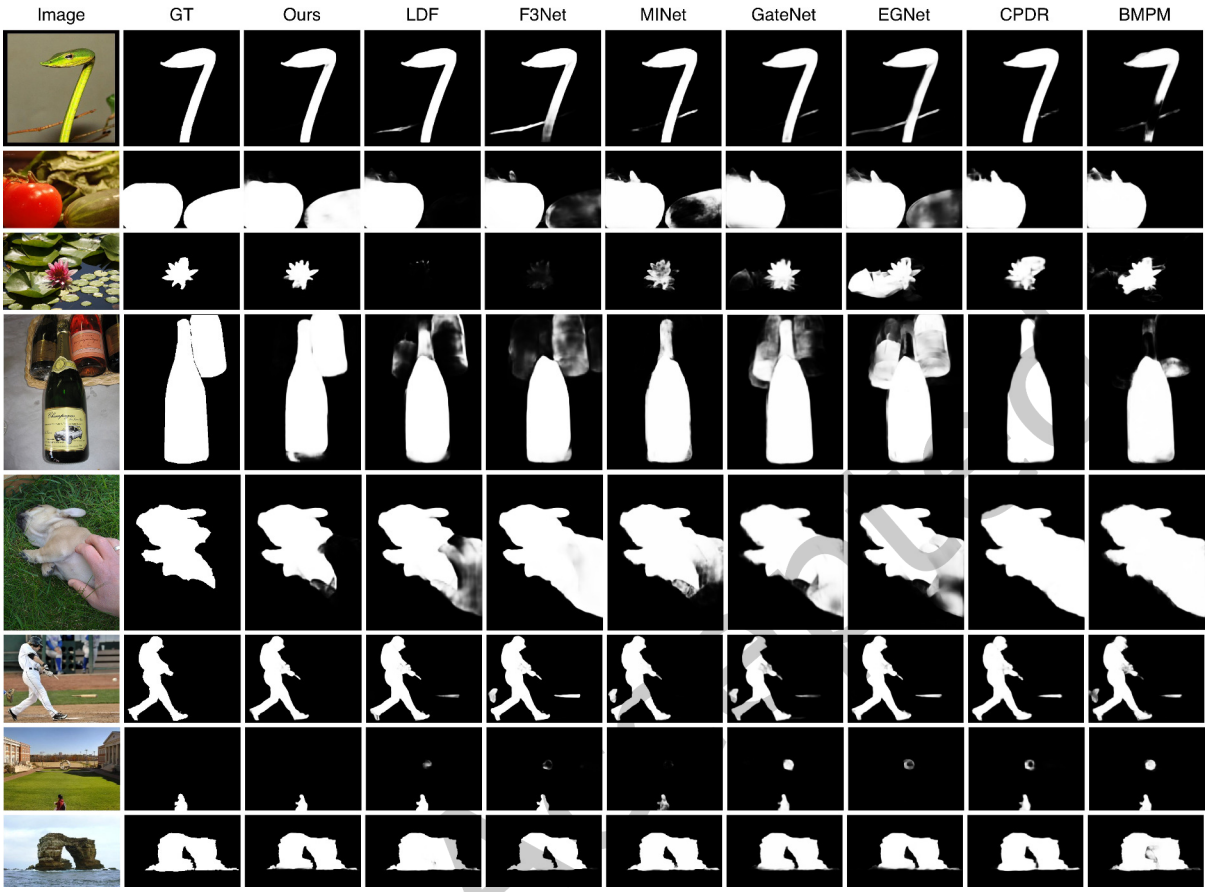

Fig. 7. Visual comparison of our method with other state-of-the-art methods in different challenging scenarios. Our method can well distinguish salient objects and suppress background noise, giving better visual appeal than others.

map to refine the boundary; 3) B3 models the detail first and then fills the body map into the detail mask; 4) B6 uses structural similarity loss to enforce the detail decoder to learn the structural information; 5) B7 adopts IoU loss to help body decoder quickly attend to body and F-loss to balance the body and detail information. 6) B4 additionally employs MDAB for better detail modeling; 7) B5 applies MBAB to fuse the detail, feature for body generation; Table 2 shows the results of the ablation study on THUR15K. As we deploy more proposed modules, the performance gains step-wise improvement, demonstrating the effectiveness of each proposed module. We then analyze the impact of each module in the following paragraphs.

Effect of Detail Modeling. We evaluate the impact of detail modeling by comparing the baseline B3 that captures detail first to baseline B1 that directly outputs the saliency map. As can be seen from Table 2, detail modeling brings about 1.5% increase in terms of both F_β and F_β^ω. Fig. 8 (*left*) illustrates the visual impact of detail modeling. We can see that the detail map can first identify the informative edges of the object and thus help the body filling part to generate saliency map with more accurate boundaries.

Impact of Generation Order. We design an interesting baseline B2 that generates intermediate body map to investigate the impact of generation order, namely taking the strategy "easier first" to let the network focus

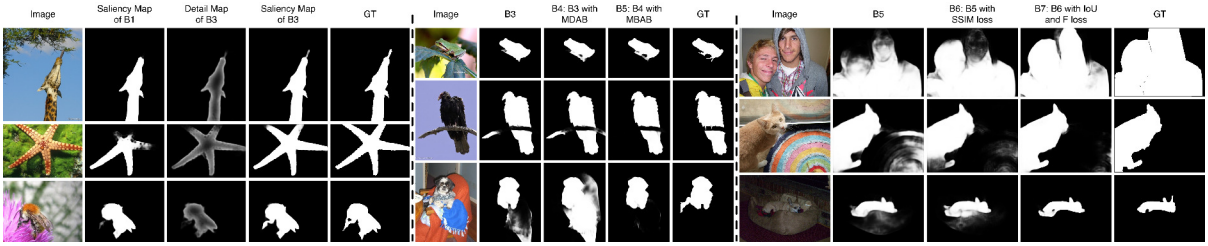

Fig. 8. (**Left**) Visual illustration of detail modeling. The first stage of B3 aims at identifying the informative details of the object, which benefits the downstream body filling task for more accurate saliency map generation. (**Middle**) Effect of multi-scale attention blocks. Equipped with the two proposed blocks, the network learns to attend to the crucial regions and remove noise to refine the mask. (**Right**) Visual effect of the hybrid loss function. l_{SSIM} improves the mask by enriching the representation with the structural information, while l_{IoU} and l_{F} helps the network to concentrate on the body and complements the body and detail map by ensuring the fused mask has a high F-score, respectively.

Table 2. Ablation studies on THUR15K.

	Setting	$MAE \downarrow$	m $F_\beta \uparrow$	$F_\beta^\omega \uparrow$
B1	Baseline	0.071	0.726	0.721
B2	B1 + Body Map → Detail Map	0.070	0.732	0.727
B3	B1 + Detail Map → Body Map	0.068	0.737	0.732
B4	B3 + l_{SSIM} on Detail Map	0.066	0.749	0.744
B5	B4+ l_{IoU} and l_{F} on Body Map	0.065	0.753	0.750
B6	B5 + MDAB	0.064	0.760	0.752
B7	B6 + MBAB	**0.062**	**0.769**	**0.755**

on the easier body map task and then refine the boundary, or the strategy "harder first" to make the model learn the harder detail first and then fill in the body. From Table 2, we can see that B3 surpasses B2, proving the effectiveness of "harder first" strategy. The reason behind may be that the neural network naturally focuses on the low-level information like the edges first then gradually shift to high-level semantics.

Effect of Multi-scale Attention Block. Based on B3 with detail modeling, we demonstrate the effect of our proposed MDAB and MBAB by setting baseline B4 and B5. Table 2 tells that the successive deployment of the two attention blocks improves the baseline by 1% in F_β and F_β^ω. As we can see from the visual illustration shown in Fig. 8 (*middle*), the two attention blocks enforce the model to concentrate on the salient object and refine the mask by removing the misclassified background region.

Effect of Hybrid Loss Function. To validate the effect of the hybrid loss function, we conduct a set of experiments over different loss configurations on our model. The evaluation results in Table 2 show that the combination of structural similarity loss on the detail map and the IoU and F loss on the body map works best. Fig. 8 (*right*) displays some visual examples to further demonstrate this. We can observe that l_{SSIM} can effectively complement the structural information and results in saliency maps with sharp and clear boundaries. On the other hand, l_{IoU} and l_{F} can ensure that the network focuses on the salient object and well combine the two maps to achieve a high F-score.

Visualizing Attention, Detail, and Body Maps. To understand how the multi-scale attention works, we visualize the local spatial-wise attention maps in the last MBAB and MDAB, the detail and body map, and the fused saliency map in Fig. 9 (*left*). As can be seen, the first detail attention map for detail decoder highlights the

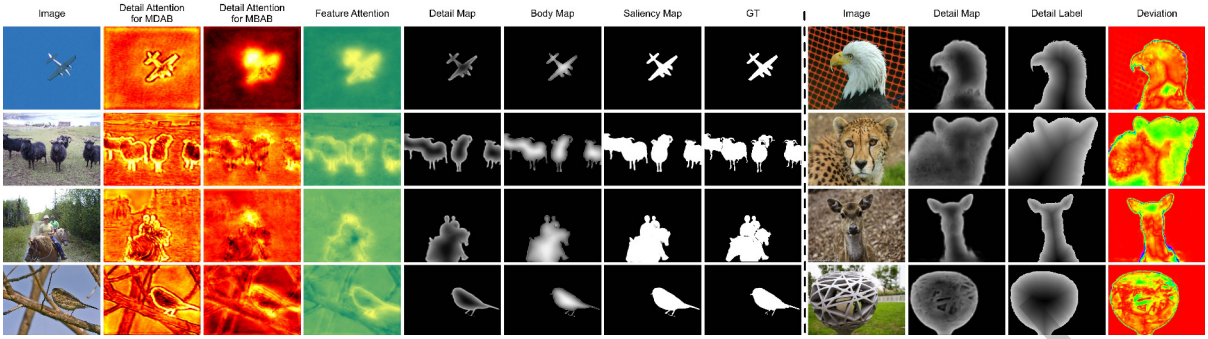

Fig. 9. (**Left**) Visual examples of the attention and output maps. (**Right**) Visualization of the deviation between predicted detail map and decomposed detail label. The detail map may also assign values to features that can characterize the object.

Fig. 10. Examples of two failure cases of our model.

edges of the object, pushing the network to concentrate on the boundary. The second detail attention for body decoder mainly emphasizes some crucial details and reveals regions that the previous attention map may neglect. The feature attention map pays attention to the important features of the image and encourages the network to segment more accurate saliency maps.

Deviation of Detail Map. To study the concrete effect of detail label, we measure the deviation between generated detail map and explicitly decoupled detail label and present some examples in Fig. 9 (*right*). We can see that the generated detail maps do not necessarily follow the exact distribution of the decoupled detail label that only assigns larger values to pixels nearby edge. Instead, the produced detail map may also assign values to crucial features that characterize the object (*e.g.*, the neck of the eagle). We expect the decoupled detail map plays the role that leads the detail map to distinguish crucial pixels based on the detail label.

Failure Cases. Fig. 10 presents two examples of failure cases on ECSSD where our model has the narrowest margin over other methods. The left example is an image that has a "smiling" ball in the center. The ground truth displays only the "smiling" ball, whereas our map predicts all the balls of high color contrast. The reason may be that our model focuses too much on the background contrast but fails to handle the inter-object contrast in this specific instance. The right example is a flower where the ground truth presents the whole flower but ours only predict its stamen. The stamen does naturally pops out of the image but is only a sub-object of the flower. We think it is because there is a lack of image-level class label supervision to let the network learns the category of an object.

5 CONCLUSION

We propose a novel end-to-end SOD framework that disentangles the original task into cascaded detail modeling and body filling. This framework can effectively reduce the difficulty of direct saliency detection. Moreover, we propose two multi-scale attention blocks that target feature fusion and help the network to generate more accurate detail and body maps. Extensive experiments have demonstrated that our method achieves state-of-the-art performances on different metrics across six datasets.

ACKNOWLEDGMENTS

This work has been supported by the EU H2020 AI4Media (No. 951911).

REFERENCES

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Süsstrunk. 2009. Frequency-tuned salient region detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

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

[3] Alessandro Bruno, Francesco Gugliuzza, Roberto Pirrone, and Edoardo Ardizzone. 2020. A Multi-Scale Colour and Keypoint Density-Based Approach for Visual Saliency Detection. *IEEE Access* 8 (2020), 121330–121343.

[4] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. 2018. Reverse attention for salient object detection. In *European Conference on Computer Vision.*

[5] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, and Shi-Min Hu. 2014. Salientshape: group saliency in image collections. *The Visual Computer* 30, 4 (2014), 443–453.

[6] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. 2014. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 3 (2014), 569–582.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *Springer International Journal of Computer Vision* 88, 2 (2010), 303–338.

[9] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. 2017. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.*

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

[11] Pedro F Felzenszwalb and Daniel P Huttenlocher. 2004. Efficient graph-based image segmentation. *Springer INTERNATIONAL JOURNAL OF COMPUTER VISION* 59, 2 (2004), 167–181.

[12] Mengyang Feng, Huchuan Lu, and Errui Ding. 2019. Attentive feedback network for boundary-aware salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[13] Chenlei Guo, Qi Ma, and Liming Zhang. 2008. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 770–778.

[15] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. 2017. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[16] Xiaodi Hou and Liqing Zhang. 2007. Saliency detection: A spectral residual approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[17] Guanbin Li and Yizhou Yu. 2016. Visual saliency detection based on multiscale deep CNN features. *IEEE Transactions on Image Processing* 25, 11 (2016), 5012–5024.

[18] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. 2014. The secrets of salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

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

[20] Nian Liu, Junwei Han, and Ming-Hsuan Yang. 2018. Picanet: Learning pixel-wise contextual attention for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[21] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[22] Zhiming Luo, Akshaya Mishra, Andrew Achkar, Justin Eichel, Shaozi Li, and Pierre-Marc Jodoin. 2017. Non-local deep features for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[23] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. 2014. How to evaluate foreground maps?. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[24] Gellért Máttyus, Wenjie Luo, and Raquel Urtasun. 2017. Deeproadmapper: Extracting road topology from aerial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.*

[25] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. 2020. Multi-Scale Interactive Network for Salient Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[26] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. 2012. Saliency filters: Contrast based filtering for salient region detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[27] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. 2019. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[28] Md Atiqur Rahman and Yang Wang. 2016. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International Symposium on Visual Computing*.
[29] Jinming Su, Jia Li, Yu Zhang, Changqun Xia, and Yonghong Tian. 2019. Selectivity or invariance: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision*.
[30] Na Tong, Huchuan Lu, Ying Zhang, and Xiang Ruan. 2015. Salient object detection via global and local cues. *Pattern Recognition* 48, 10 (2015), 3258–3267.
[31] Anne M Treisman and Garry Gelade. 1980. A feature-integration theory of attention. *Cognitive Psychology* 12, 1 (1980), 97–136.
[32] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. 2017. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[33] Tiantian Wang, Lihe Zhang, Shuo Wang, Huchuan Lu, Gang Yang, Xiang Ruan, and Ali Borji. 2018. Detect globally, refine locally: A novel approach to saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[34] Wenguan Wang, Qiuxia Lai, Huazhu Fu, Jianbing Shen, Haibin Ling, and Ruigang Yang. 2021. Salient object detection in the deep learning era: An in-depth survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).
[35] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. 2003. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*.
[36] Jun Wei, Shuhui Wang, and Qingming Huang. 2020. F^3Net: Fusion, Feedback and Focus for Salient Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
[37] Jun Wei, Shuhui Wang, Zhe Wu, Chi Su, Qingming Huang, and Qi Tian. 2020. Label Decoupling Framework for Salient Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[38] Runmin Wu, Mengyang Feng, Wenlong Guan, Dong Wang, Huchuan Lu, and Errui Ding. 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*.
[39] Zhe Wu, Li Su, and Qingming Huang. 2019. Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[40] Zhe Wu, Li Su, and Qingming Huang. 2019. Stacked cross refinement network for edge-aware salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
[41] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. 2013. Hierarchical saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[42] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. 2013. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[43] Lu Zhang, Ju Dai, Huchuan Lu, You He, and Gang Wang. 2018. A bi-directional message passing model for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[44] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. 2017. Amulet: Aggregating multi-level convolutional features for salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
[45] Qing Zhang, Jiajun Lin, Yanyun Tao, Wenju Li, and Yanjiao Shi. 2017. Salient object detection via color and texture cues. *Neurocomputing* 243 (2017), 35–48.
[46] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. 2018. Progressive attention guided recurrent network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[47] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. 2019. EGNet: Edge guidance network for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision*.
[48] Kai Zhao, Shanghua Gao, Wenguan Wang, and Ming-Ming Cheng. 2019. Optimizing the f-measure for threshold-free salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
[49] Ting Zhao and Xiangqian Wu. 2019. Pyramid feature attention network for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[50] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. 2020. Suppress and balance: A simple gated network for salient object detection. In *European Conference on Computer Vision*.
[51] Yijie Zhong, Bo Li, Lv Tang, Hao Tang, and Shouhong Ding. 2021. Highly Efficient Natural Image Matting. In *British Machine Vision Conference*.