Selectivity or Invariance: Boundary-aware Salient Object Detection

Jinming Su[1,3], Jia Li[1,3*], Yu Zhang[1], Changqun Xia[3] and Yonghong Tian[2,3*]
[1]State Key Laboratory of Virtual Reality Technology and Systems, SCSE, Beihang University
[2]National Engineering Laboratory for Video Technology, School of EE&CS, Peking University
[3]Peng Cheng Laboratory, Shenzhen, China
{sujm, jiali}@buaa.edu.cn, zhangyulb@gmail.com, xiachq@pcl.ac.cn, yhtian@pku.edu.cn

Abstract

Typically, a salient object detection (SOD) model faces opposite requirements in processing object interiors and boundaries. The features of interiors should be invariant to strong appearance change so as to pop-out the salient object as a whole, while the features of boundaries should be selective to slight appearance change to distinguish salient objects and background. To address this selectivity-invariance dilemma, we propose a novel boundary-aware network with successive dilation for image-based SOD. In this network, the feature selectivity at boundaries is enhanced by incorporating a boundary localization stream, while the feature invariance at interiors is guaranteed with a complex interior perception stream. Moreover, a transition compensation stream is adopted to amend the probable failures in transitional regions between interiors and boundaries. In particular, an integrated successive dilation module is proposed to enhance the feature invariance at interiors and transitional regions. Extensive experiments on six datasets show that the proposed approach outperforms 16 state-of-the-art methods.

1. Introduction

Salient object detection (SOD), which aims to detect and segment objects that can capture human visual attention, is an important step before subsequent vision tasks such as object recognition [31], tracking [13] and image parsing [18]. To address the SOD problem, hundreds of learning-based models [15, 36, 39, 7, 48, 11] have been proposed in the past decades, among which the state-of-the-art deep models [36, 7] have demonstrated impressive performance on many datasets [42, 43, 24, 21, 34, 39, 10]. However, there still exist two key issues that need to be further addressed. First, the interiors of a large salient object may have large appearance change, making it difficult to detect the salient object as a whole (see Fig. 1(a)(b)). Second, the boundaries of salient objects may be very weak so that they cannot be distinguished from the surrounding background regions (see Fig. 1(c)(d)). Due to these two issues, SOD remains a challenging task even in the deep learning era.

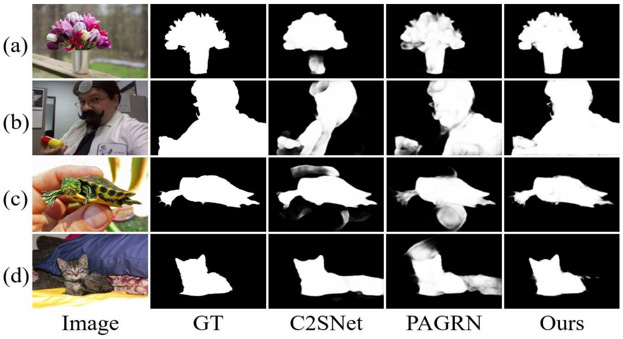


Figure 1. Different regions of salient objects require different features. (a)(b) Features of interiors should be invariant to large appearance change to detect the salient object as a whole; (c)(d) features at boundaries should be selective to distinguish the slight differences between salient objects and background regions. Images and ground-truth masks are from [42]. Results are generated by C2SNet [23], PAGRN [50] and our approach.

By further investigating these two issues at object interiors and boundaries, we find that the challenge may be mainly from the selectivity-invariance dilemma [19]. In the interiors, the features extracted by a SOD model should be invariant to various appearance changes such as size, color and texture. Such invariant features ensure that the salient object can pop-out as a whole. However, the features at boundaries should be sufficiently selective at the same time so that the minor difference between salient objects and background regions can be well distinguished. In other words, different regions of a salient object poses different requirements for a SOD model, and such dilemma actually prevents the perfect segmentation of salient objects with various sizes, appearances and contexts.

[*] Correspondence should be addressed to Jia Li and Yonghong Tian. The code and models are available on http://cvteam.net.

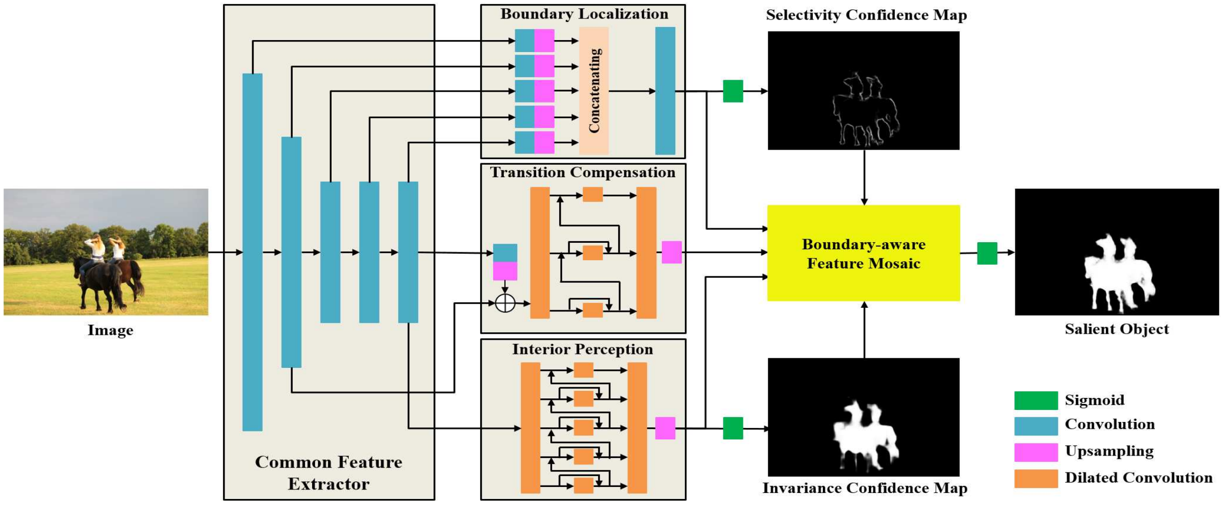


Figure 2. The framework of our approach. We first use ResNet-50 to extract common features for three streams. The boundary localization stream uses multi-level features and a simple network to detect salient boundaries with high selectivity, while the interior perception stream uses single-level features and a complex network to guarantee invariance in salient interiors. Their output features are used to form two confidence maps of selectivity and invariance, based on which a transition compensation stream is adopted to amend the probable failures that are likely to occur in the transition regions between boundaries and interiors. These three streams are concatenated to form a boundary-aware feature mosaic map so that the salient object can pop-out as a whole with clear boundaries.

To break out of this dilemma, a feasible solution is to adopt different feature extraction strategies at object interiors and boundaries. Inspired by that, we propose a boundary-aware network with successive dilation for image-based SOD. As shown in Fig. 2, the network first extracts common visual features and then deliver them into three separate streams. Among these three streams, the boundary localization stream is a simple subnetwork that aims to extract selective features for detecting the boundaries of salient objects, while the interior perception stream emphasizes the feature invariance in detecting the salient objects. In addition, a transition compensation stream is adopted to amend the probable failures that may occur in the transitional regions between interiors and boundaries, where the feature requirement gradually changes from invariance to selectivity. Moreover, an integrated successive dilation module is proposed to enhance the capability of the interior perception and transition compensation streams so that they can extract invariant features for various visual patterns. Finally, the output of these three streams are adaptively fused to generate the masks of salient objects. Experimental results on six public benchmark datasets show that our approach outperforms 16 state-of-the-art SOD models. Moreover, our approach demonstrates impressive capability in accurately segmenting salient boundaries at fine scales.

The main contributions of this paper include: 1) we revisit the problem of SOD from the perspective of selectivity-invariance dilemma, which may be helpful to develop new models; 2) we propose a novel boundary-aware network for salient object detection, which consistently outperforms 16 state-of-the-art algorithms on six datasets; 3) we propose an integrated successive dilation module that can enhance the capability of extracting invariant features.

2. Related Work

Hundreds of image-based SOD methods have been proposed in the past decades. Early methods mainly adopted hand-crafted local and global visual features as well as heuristic saliency priors such as the color difference [1], distance transformation [33] and local/global contrast [16, 8]. More details about the traditional methods can be found in the survey [2]. In this review, we mainly focus on the latest deep models in recent three years.

Lots of these deep models are devoted to fully utilizing the feature integration to enhance the performance of neural networks [20, 22, 25, 36, 48, 29, 49, 50, 47]. For example, Zhang *et al.* [50] proposed an attention guided network to selectively integrates multi-level information in a progressive manner. Wang *et al.* [36] proposed a pyramid pooling module and a multi-stage refinement mechanism to gather contextual information and stage-wise results, respectively. Zhang *et al.* [48] adopted a framework to aggregate multi-level convolutional features into multiple resolutions, which were then combined to predict saliency maps in a recursive manner. Luo *et al.* [29] proposed a simplified convolutional neural network by combining global and local information through a multi-resolution 4×5 grid structure. Zhang *et al.* [49] utilized the deep uncertain convolutional features and proposed a reformulated dropout after specific convolutional layers to construct an uncertain ensemble of internal feature units. Different with them, we propose an integrated successive dilation module to capture richer contextual information to produce features that account for interior invariance and introduce skip connections from low-level features to promote selective representations of boundaries.

In addition, many models [37, 7, 3, 6, 46, 38, 23] comprehend saliency detection task by relating other vision tasks. Chen *et al.* [6] proposed reverse attention mechanism which is inspired from human perception process by using top information to guide bottom-up feed-forward process in a top-down manner. Chen *et al.* [7] incorporated human fixation with semantic information to simulate the human annotation process for salient objects. Chen and Li [3] proposed a complementarity-aware network to fuse both cross-model and cross-level features to solve saliency detection task with depth information. Wang *et al.* [38] proposed to learn the local contextual information for each spatial position to refine boundaries. Li *et al.* [23] considered contours as useful priors and proposed to facilitate feature learning in SOD by transferring knowledge from an existing contour detection model. Our work differs with them by fusing the boundary and interior features of salient objects with a compensation mechanism and an adaptive manner.

3. The Proposed Approach

The selectivity-invariance dilemma in SOD indicates that the boundaries and interiors of salient objects require different types of features. Inspired by that, we propose a boundary-aware network for saliency detection (see Fig. 2 for the system framework). The network first extract common features, which are then processed with three separate streams. The outputs of these streams are then fused to generate the final masks of salient objects in a boundary-aware feature mosaic selection manner. Details of the proposed approach are descried as follows.

3.1. Common Feature Extraction

As shown in Fig. 2, the boundary-aware network starts with ResNet-50 [12]. As a common feature extractor, we remove the last global pooling and fully connected layers and use only the five residual blocks. For the sake of simplification, the subnetworks in these five blocks are denoted as $\theta_i(\pi_i), i \in \{1, \dots, 5\}$, where π_i is the set of parameters of θ_i. Note that the input of $\theta_i(\pi_i)$ is the output of $\theta_{i-1}(\pi_{i-1}), \forall\ i = 2, \dots, 5$, and we omit the input for the sake of simplification. In addition, the strides of all convolutional layers in θ_4 and θ_5 are set to 1 to avoid the over downsample of feature maps. As in [45], we enlarge the receptive fields by using the dilation of 2 and 4 in all convolutional layers of θ_4 and θ_5, respectively. Finally, for a $H \times W$ input image, the subnetwork θ_5 outputs a $\frac{H}{8} \times \frac{W}{8}$ feature map with 2048 channels.

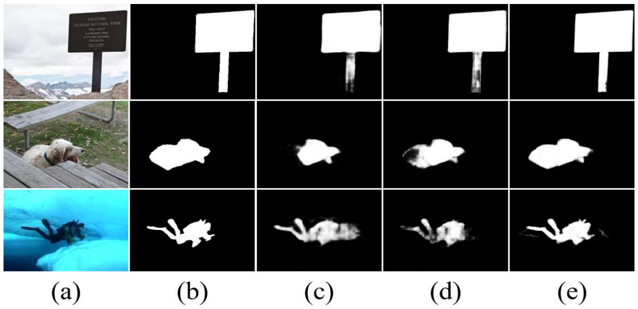


Figure 3. The SOD results from the combination of three streams. (a) Image; (b) ground-truth; (c) only interior perception stream; (d) interior perception and boundary localization streams; (e) three streams. We can see that interior perception stream may fail near object boundaries due to the emphasis of invariance, while such vague boundaries can be corrected by incorporating the boundary localization stream with the emphasis of selectivity. In addition, the probable failures of these two streams can be amended by the transition compensation stream.

3.2. Boundary-aware SOD with Three Streams

Given the common features, we use three streams for boundary localization, interior perception and transition compensation, respectively. The boundary stream is inspired by the work of [41], which is a simple subnetwork $\phi_{\mathcal{B}}(\pi_{\mathcal{B}})$ that aggregates multi-level common features and fuses them by upsampling and concatenating to obtain the final boundary predictions. The input of this subnetwork is the concatenation of features from $\{\theta_i(\pi_i)\}_{i=1}^{5}$. For the feature map of each $\theta_i(\pi_i)$, we add two convolution layers with 128 kernels of 3×3 and one 1×1 kernel, respectively. These two layers are used to squeeze the common features, which are then upsampled to $H \times W$. After the concatenation, we add an extra layer with one 1×1 kernel to output a single channel $H \times W$ feature map $\phi_{\mathcal{B}}(\pi_{\mathcal{B}})$. A sigmoid layer is then used to generate a selectivity confidence map that is expected to approximate the boundary map of salient objects (denoted as $G_{\mathcal{B}}$) by minimizing the loss

$$L_{\mathcal{B}} = E(Sig(\phi_{\mathcal{B}}(\pi_{\mathcal{B}})), G_{\mathcal{B}}), \tag{1}$$

where $Sig(\cdot)$ is the sigmoid function and $E(\cdot, \cdot)$ means the cross-entropy loss function. By taking multi-level features as the input and using only simple feature mapping subnetworks, the boundary localization stream demonstrates a strong selectivity at object boundaries.

Different from the boundary localization stream, the interior perception stream $\phi_{\mathcal{I}}(\pi_{\mathcal{I}})$ emphasizes feature invariance inside large salient objects. Therefore, it takes less input features and uses a more complex subnetwork. Its input is the output of the last common feature extractor $\theta_5(\pi_5)$, and the output is a single-channel $H \times W$ feature map $\phi_{\mathcal{I}}(\pi_{\mathcal{I}})$. Similarly, we can use the sigmoid operation to derive an invariance confidence map, which is expected to approximate the ground-truth mask of salient objects G by minimizing the cross-entropy loss:

$$L_{\mathcal{I}} = E(Sig(\phi_{\mathcal{I}}(\pi_{\mathcal{I}})), G), \tag{2}$$

Note that an integrated successive dilation (ISD) module is used in this stream to map the input to the output by perceiving local contexts at successive scales, which will be introduced in the next subsection.

As shown in Fig. 3, the awareness of boundaries can be enhanced by handling the boundaries and interior regions with two separate streams: one uses multi-level features and a simple network to emphasize selectivity, the other one uses single-level features and a complex network to enhance invariance. However, the combination of these two streams may still have failures, especially for the transitional regions between interiors and boundaries that require a balance of selectivity and invariance. To this end, we adopt a transition compensation stream $\phi_{\mathcal{T}}(\pi_{\mathcal{T}})$ to adaptively amend these failures by compensating features in the transitional regions. Different from the first two streams, $\phi_{\mathcal{T}}(\pi_{\mathcal{T}})$ takes the element-wise summation of the two-level features (one high-level $\theta_5(\pi_5)$ and one low-level $\theta_2(\pi_2)$) as the input. In this manner, localization-aware fine-level features and semantic-aware coarse-level ones can jointly enrich the representation power within transition regions. Since $\theta_2(\pi_2)$ has the resolution $\frac{H}{4} \times \frac{W}{4}$, we upsample $\theta_5(\pi_5)$ to $\frac{H}{4} \times \frac{W}{4}$ after two pre-processing layers using 256 kernels of 3×3 and 256 kernel of 1×1, respectively. Based on these features, an ISD module with medium complexity is used to generate a transitional feature representation map that mediates both selectivity and invariance, which ensures that detailed structures of salient objects to be correctly detected.

Instead of approximating certain "ground-truth", the parameters of the transition compensation stream are supervised by the feedback from both the ground-truth masks of salient objects and the predictions of boundary and interior streams. Suppose that this stream also outputs a $H \times W$ feature map $\phi_{\mathcal{T}}$ after upsampling, we combine it with the feature maps $\phi_{\mathcal{B}}$ and $\phi_{\mathcal{I}}$. Note that features in $\phi_{\mathcal{B}}$, $\phi_{\mathcal{I}}$ and $\phi_{\mathcal{T}}$ emphasize selectivity, invariance and their tradeoff, respectively. As a result, the direct element-wise summation or concatenation may incorporate unexpected noises as shown in Fig. 4. To reduce these noises, we adopt a boundary-aware mosaic approach that assigns different strengths to different regions, guided by confidence confidence maps from boundary and interior streams. This approach ensures a well-learned combination of $\phi_{\mathcal{T}}$ and $\phi_{\mathcal{B}}$ as well as $\phi_{\mathcal{I}}$ by properly balancing selectivity and invariance. Let $\mathbf{M}$ be the feature mosaic map, we combine the three maps according to the selectivity confidence map $\mathbf{M}_{\mathcal{B}} = Sig(\phi_{\mathcal{B}})$ and the invariance confidence map $\mathbf{M}_{\mathcal{I}} = Sig(\phi_{\mathcal{I}})$:

$$\begin{aligned}\mathbf{M} =& \phi_{\mathcal{B}} \otimes (1 - \mathbf{M}_{\mathcal{I}}) \otimes \mathbf{M}_{\mathcal{B}} + \phi_{\mathcal{I}} \otimes \mathbf{M}_{\mathcal{I}} \otimes (1 - \mathbf{M}_{\mathcal{B}}) \\ &+ \phi_{\mathcal{T}} \otimes (1 - \mathbf{M}_{\mathcal{I}}) \otimes (1 - \mathbf{M}_{\mathcal{B}}),\end{aligned} \tag{3}$$

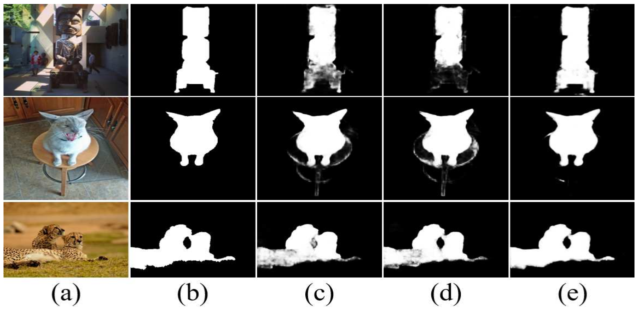


Figure 4. Results of different combinations of three streams $\phi_{\mathcal{B}}$, $\phi_{\mathcal{I}}$ and $\phi_{\mathcal{T}}$. (a) Image; (b) ground-truth; (c) element-wise summation; (d) concatenation; (e) our mosaic approach.

where $\otimes$ denotes the element-wise product between two matrices. We can see that the first term emphasizes selective features $\phi_{\mathcal{B}}$ at the locations with high selectivity and low invariance confidences, while the second term emphasizes invariant features $\phi_{\mathcal{I}}$ at the locations with high invariance and low selectivity confidence. For the other locations with medium selectivity and invariance confidences, the transitional features $\phi_{\mathcal{T}}$ will be added to correct the features. In other words, the transition stream actually learns to approximate the uncertain regions in the other two streams by providing feature compensations. After that, we can derive final saliency map as $Sig(\mathbf{M})$ by minimizing the loss

$$L_0 = E(Sig(\mathbf{M}), G), \tag{4}$$

which indirectly supervises the training process of $\phi_{\mathcal{T}}(\pi_{\mathcal{T}})$. By taking the losses of Eqs. (1), (2) and (4), the overall learning objective can be formulated as

$$\min_{\{\pi_i\}_{i=1}^5, \pi_{\mathcal{B}}, \pi_{\mathcal{I}}, \pi_{\mathcal{T}}} L_0 + L_{\mathcal{B}} + L_{\mathcal{I}}. \tag{5}$$

From Eq. (5), we can see that the parameters $\{\pi_i\}_{i=1}^5$ and $\pi_{\mathcal{T}}$ are supervised by the three losses, while the parameters $\pi_{\mathcal{B}}, \pi_{\mathcal{I}}$ are supervised by two losses. Note that the boundary information is used in L_0, $L_{\mathcal{B}}$, and the generation of the feature mosaic map $\mathbf{M}$ is also guided by the selectivity confidence map, making the whole network aware of object boundaries.

3.3. Integrated Successive Dilation Module

In the interior perception stream and the transition compensation streams, the key requirement is to extract invariant features for a region embedded in various contexts. To enhance such capability, we propose an integrated successive dilation module (named as ISD) to efficiently aggregate contextual information at a sequence of scales for the purpose of enhancing the feature invariance.

The ISD module with N parallel branches with skip connections is denoted as ISD-N, and we show the structure of ISD-5 in Fig. 5 as an example. The first layer of each

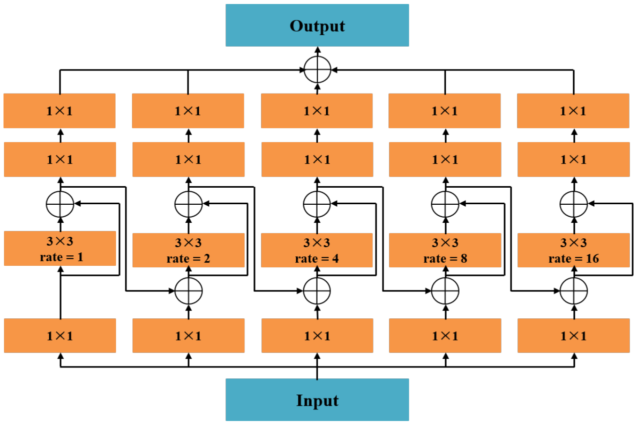


Figure 5. Structure of the integrated successive dilation (ISD) module. 1 $\times 1$ and 3 $\times 3$ means the convolutional kernel size, and rate represents the dilation rate in dilated convolution.

branch is a convolutional layer with 1×1 kernels that is used for channel compression. The second layer of each branch adopts dilated convolution, in which the dilation rates start from 1 in the first branch and double in the subsequent branch. In this manner, the last branch has a dilation rate of 2^{N-1}. By adding intra- and inter-branch short connections, the feature map generated by a branch layer actually integrates the perception results of the previous branch and further handle them with larger dilation. In this way, the feature map from the first branch of the second layer is also encoded in the feature maps of subsequent branches, which actually gets processed by successive dilation rates. In other words, an ISD-N module gains the capability of perceiving various local contexts with the smallest dilation rate of 1 and the largest dilation rate of $2^N - 1$. After that, the third and the forth layers adopt 1×1 kernels to integrate feature maps formed under various dilation rates. In practice, we use ISD-5 in the interior perception stream and ISD-3 in the transition compensation streams.

4. Experiments and Results

4.1. Experimental Setup

Datasets. To evaluate the performance of the proposed approach, we conduct experiments on six benchmark datasets [42, 43, 24, 21, 34, 39]. Details of these datasets are described briefly as follows: ECSSD [42] contains 1,000 images with complex structures and obvious semantically meaningful objects. DUT-OMRON [43] consists of 5,168 complex images with pixel-wise annotations of salient objects and all images are downsampled to a maximal side length of 400 pixels. PASCAL-S [24] includes 850 natural images that are pre-segmented into objects or regions and free-viewed by 8 subjects in eye-tracking tests for salient object annotation. HKU-IS [21] comprises 4,447 images and lots of images contain multiple disconnected salient objects or salient objects that touch image boundaries. DUTS [34] is a large scale dataset containing 10533 training images (denoted as DUTS-TR) and 5019 test images(denoted as DUTS-TE). The images are challenging with salient objects that occupy various locations and scales as well as complex background. XPIE [39] has 10000 images covering a variety of simple and complex scenes with salient objects of different numbers, sizes and positions.

Evaluation Metrics. We adopt mean absolute error (MAE), F-measure (F_β) score, weighted F-measure (F_β^w) score [30], Precision-Recall (PR) curve and F-measure curve as our evaluation metrics. MAE reflects the average pixel-wise absolute difference between the estimated and ground-truth saliency maps. In computing F_β, we normalize the predicted saliency maps into the range of [0, 255] and binarize the saliency maps with a threshold sliding from 0 to 255 to compare the binary maps with ground-truth maps. At each threshold, Precision and Recall can be computed. F_β is computed as:

$$F_\beta = \frac{(1+\beta^2) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}. \tag{6}$$

where we set $\beta^2 = 0.3$ to emphasize more on Precision than Recall as suggested in [1]. Then we can plot the PR curve and F-measure curve based on all the binary maps over all saliency maps in a given dataset.

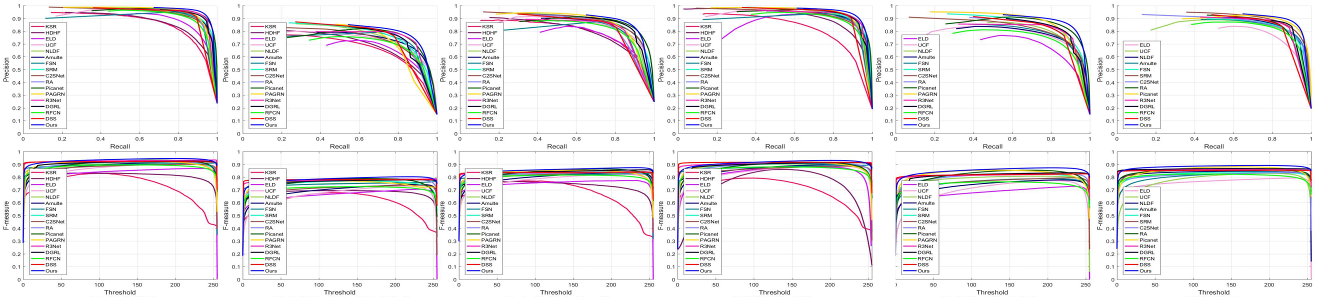
We report F_β using an adaptive threshold for generating a binary saliency map and the threshold is computed as twice the mean of a saliency map. Besides, F_β^w is computed to reflect the overall performance (refer to [30] for details).

Training and Inference. We use standard stochastic gradient descent algorithm to train the whole network end-to-end with the cross-entropy losses between estimated and ground-truth maps. In the optimization process, the parameter of common feature extractor is initialized by the pre-trained ResNet-50 model [12], whose learning rate is set to 5×10^{-9} with a weight decay of 0.0005 and momentum of 0.9. The learning rates of the rest layers in our network are set to 10 times larger. Besides, we employ the "poly" learning rate policy for all experiments similar to [28].

We train our network on DUTS-TR [34] as used in [38, 26, 36]. For a more comprehensive demonstration, we also trained our network with VGG-16 [32] on MSRA10K [8] as used in [49, 48, 7, 23] and on DUTS-TR as done in [50, 26]. The training images are not done with any special treatment except the horizontal flipping. The training process takes about 15 hours and converges after 200k iterations with mini-batch of size 1. During testing, the proposed network removes all the losses, and each image is directly fed into the network to obtain its saliency map without any pre-processing. The proposed method runs at about 13 fps with about 400×300 resolution on our computer with a 3.60GHz CPU and a GTX 1080ti GPU.

Table 1. Performance of 16 state-of-the-arts and the proposed method on six benchmark datasets. Smaller MAE, larger F_β^w and F_β correspond to better performance. The best results of different backbones are in **blue** and **red** fonts. "-" means the results cannot be obtained and "†" means the results are post-processed by dense conditional random field (CRF) [17]. Note that the backbone of PAGRN is VGG-19 [32] and the one of R3Net is ResNeXt-101 [40]. MK: MSRA10K [8], DUTS: DUTS-TR [34], MB: MSRA-B [27].

Models	Training dataset	ECSSD			DUT-OMRON			PASCAL-S			HKU-IS			DUTS-TE			XPIE		
		MAE	F_β^w	F_β	MAE	F_β^w	F_β	MAE	F_β^w	F_β	MAE	F_β^w	F_β	MAE	F_β^w	F_β	MAE	F_β^w	F_β
VGG-16 backbone [32]																			
KSR [37]	MB	0.132	0.633	0.810	0.131	0.486	0.625	0.157	0.569	0.773	0.120	0.586	0.773	-	-	-	-	-	-
HDHF [22]	MB	0.105	0.705	0.834	0.092	0.565	0.681	0.147	0.586	0.761	0.129	0.564	0.812	-	-	-	-	-	-
ELD [20]	MK	0.078	0.786	0.829	0.091	0.596	0.636	0.124	0.669	0.746	0.063	0.780	0.827	0.092	0.608	0.647	0.085	0.698	0.746
UCF [49]	MK	0.069	0.807	0.865	0.120	0.574	0.649	0.116	0.696	0.776	0.062	0.779	0.838	0.112	0.596	0.670	0.095	0.693	0.773
NLDF [29]	MB	0.063	0.839	0.892	0.080	0.634	0.715	0.101	0.737	0.806	0.048	0.838	0.884	0.065	0.710	0.762	0.068	0.762	0.825
Amulet [48]	MK	0.059	0.840	0.882	0.098	0.626	0.673	0.099	0.736	0.795	0.051	0.817	0.853	0.085	0.658	0.705	0.074	0.743	0.796
FSN [7]	MK	0.053	0.862	0.889	0.066	0.694	0.733	0.095	0.751	0.804	0.044	0.845	0.869	0.069	0.692	0.728	0.066	0.762	0.812
C2SNet [23]	MK	0.057	0.844	0.878	0.079	0.643	0.693	0.086	0.764	0.805	0.050	0.823	0.854	0.065	0.705	0.740	0.066	0.764	0.807
RA [6]	MB	0.056	0.857	0.901	0.062	0.695	0.736	0.105	0.734	0.811	0.045	0.843	0.881	0.059	0.740	0.772	0.067	0.776	0.836
Picanet [26]	DUTS	0.046	0.865	0.899	0.068	0.691	0.730	**0.079**	0.775	0.821	0.042	0.847	0.878	0.054	0.747	0.770	0.053	0.799	0.841
PAGRN [50]	DUTS	0.061	0.834	0.912	0.071	0.622	0.740	0.094	0.733	0.831	0.048	0.820	0.896	0.055	0.724	0.804	-	-	-
RFCN [35]	MK	0.067	0.824	0.883	0.077	0.635	0.700	0.106	0.720	0.802	0.055	0.803	0.864	0.074	0.663	0.731	0.073	0.736	0.809
DSS† [14]	MB	0.052	0.872	**0.918**	0.063	0.697	**0.775**	0.098	0.756	0.833	0.040	0.867	**0.904**	0.056	0.755	**0.810**	0.065	0.784	0.849
BANet	MK	0.046	0.873	0.907	0.062	0.705	0.742	0.082	0.780	0.832	0.041	0.851	0.883	0.048	0.766	0.791	0.052	0.808	0.853
BANet	DUTS	**0.041**	**0.890**	0.917	**0.061**	**0.719**	0.750	**0.079**	**0.794**	**0.839**	**0.037**	**0.869**	0.893	**0.046**	**0.781**	0.805	**0.048**	**0.822**	**0.862**
ResNet-50 backbone [12]																			
SRM [36]	DUTS	0.054	0.853	0.902	0.069	0.658	0.727	0.086	0.759	0.820	0.046	0.835	0.882	0.059	0.722	0.771	0.057	0.783	0.841
Picanet [26]	DUTS	0.047	0.866	0.902	0.065	0.695	0.736	0.077	0.778	0.826	0.043	0.840	0.878	0.051	0.755	0.778	0.052	0.799	0.843
DGRL [38]	DUTS	0.043	0.883	0.910	0.063	0.697	0.730	0.076	0.788	0.826	0.037	0.865	0.888	0.051	0.760	0.781	0.048	0.818	0.859
R3† [9]	MK	0.040	0.902	0.924	0.063	0.728	**0.768**	0.095	0.760	0.834	0.036	0.877	0.902	0.057	0.765	0.805	0.058	0.805	0.854
BANet	DUTS	**0.035**	**0.908**	**0.929**	**0.059**	**0.736**	0.763	**0.072**	**0.810**	**0.849**	**0.032**	**0.886**	**0.905**	**0.040**	**0.811**	**0.829**	**0.044**	**0.839**	**0.873**



ECSSD DUT-OMRON PASCAL-S HKU-IS DUTS-TE XPIE

Figure 6. The PR curves and F-measure curves of 16 state-of-the-arts and our approach are listed across six benchmark datasets.

4.2. Comparisons with the State-of-the-Arts

We compare our approach denoted as **BANet** with 16 state-of-the-art methods, including KSR [37], HDHF [22], ELD [20], UCF [49], NLDF [29], Amulet [48], FSN [7], SRM [36], C2SNet [23], RA [6], Picanet [26], PAGRN [50], R3Net [9], DGRL [38], RFCN [35] and DSS [14]. For fair comparison, we obtain the saliency maps of these methods from authors or the deployment codes provided by authors.

Quantitative Evaluation. The proposed approach is compared with 16 state-of-the-art saliency detection methods on six datasets. The comparison results are shown in Tab.1 and Fig. 6. From Tab.1, we can see that our method consistently outperforms other methods across all the six benchmark datasets. It is worth noting that F_β^w of our method is significantly better compared with the second best results on PASCAL-S (0.810 against 0.788), DUTS-TE (0.811 against 0.765) and XPIE (0.839 against 0.818), and have similar improvements on the other datasets. F_β also has obvious improvement on all the datasets except DUT-OMRON, on which we achieve the third but the best two methods both employ dense CRF [17] to further refine their results. As for MAE, our approach also achieves the best performance on all the datasets. For overall comparisons, PR and F-measure curve of different methods are displayed in Fig. 6. One can observe that our approach noticeably outperforms all the other methods. These observations demonstrate the efficiency of boundary-aware net-

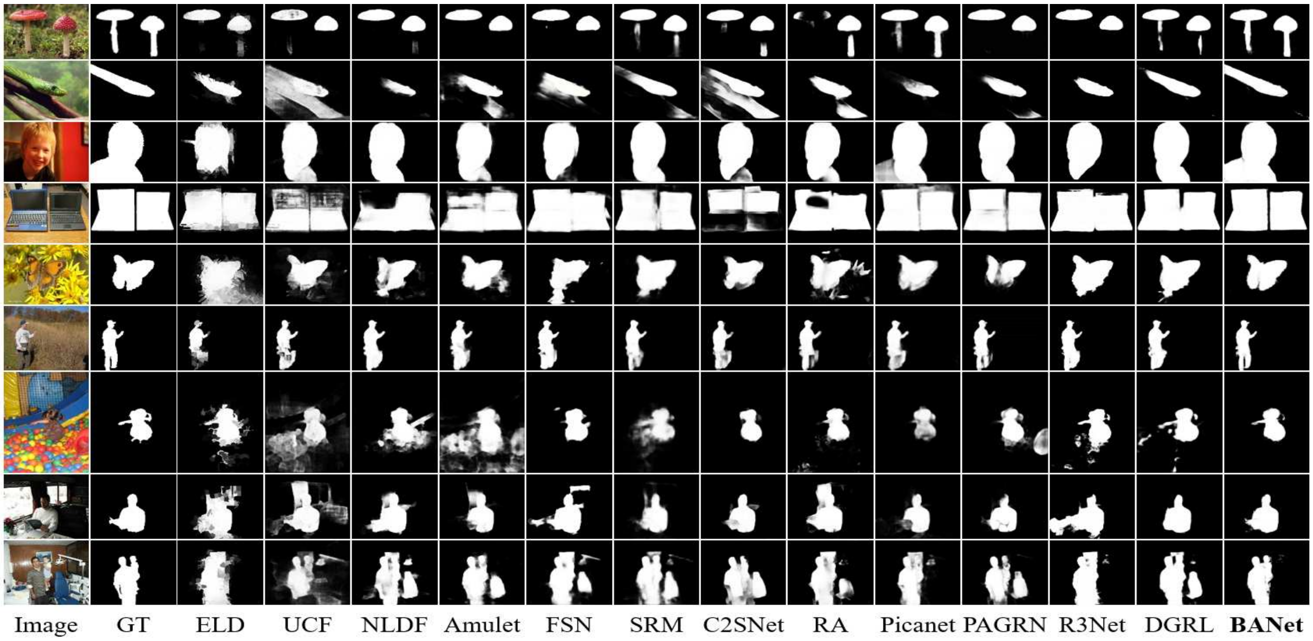


Figure 7. Qualitative comparisons of the state-of-the-art algorithms and our approach. GT means ground-truth masks of salient objects. The images are selected from six datasets for testing.

work, which indicates that it is useful to deal with the problem of SOD from the perspective of selectivity-invariance dilemma. Note that the results of DSS, RA and HDHF on HKU-IS [21] are only conducted on the test set.

Qualitative Evaluation. Fig. 7 show examples of saliency maps generated by our approach as well as other state-of-the-art methods. We can see that salient objects can pop-out as a whole with clear boundaries by the proposed method. From Fig. 7, we can find that many methods fail to detect the salient objects with large changed appearance as a whole as depicted in the row of 1 to 3. This indicates the feature invariance is important for SOD, which can be extracted by ISD to guarantee the integrity of salient objects. In addition, when salient objects share the same attributes (such as color, texture and locations) with background, the boundaries of salient objects predicted by many methods become vague, as shown in the row of 4 to 6. In our approach, the feature selectivity at boundary is guaranteed by the awareness of boundaries, which deal with the above situation to obtain clear boundaries. Moreover, three extra examples about more difficult scenes are shown in the last three rows of Fig. 7, our methods also obtain the impressive results. These observations indicated the feature selectivity and invariance are important to deal with the integrity of objects and clarity of boundaries for SOD.

4.3. Ablation Analysis

To validate the effectiveness of different components of the proposed method, we conduct several experiments on all the six datasets to compare the performance variations of our methods with different experimental settings.

Effectiveness of the BLS and TCS. To investigate the efficacy of the proposed boundary localization stream (BLS) and transition compensation stream (TCS), we conduct ablation experiments across all six datasets by introducing two different settings for comparisons. The first setting denoted as “IPS” contains only the interior perception stream following the common feature extractor to directly predict the saliency maps. To explore the effectiveness of boundary localization stream, the second one named as “IPS + BLS” utilizes the interior perception stream and boundary localization stream together, where the final predicted results are added up directly without the transition compensation stream. Note that our proposed approach **BANet** combines all three streams.

For a comprehensive comparison, above-mentioned settings and BANet are evaluated on six benchmark datasets. The comparison results are listed in Tab. 2. We can observe that although only BLS is utilized compared with IPS, the MAE obviously decreases and F-measure scores increase significantly as shown in the second row. This indicates that the BLS provides a strong selectivity at object boundaries that boosts the performance a remarkable improvement. Besides, combined with TCS on the basis of the second setting, the performance of the model is further improved by amending the probable failures in transitional regions between boundaries and interiors. For example, the F_β score increases from 0.914 to 0.929, with an improvement up to 1.5% on HKU-IS dataset. We also provide examples of different settings. As shown in Fig. 3, with the cooperation of

Table 2. Performance of the three streams in the proposed approach on six benchmark datasets. IP means interior perception stream, "IP + BL" means the combination of interior perception and boundary localization streams, and BANet is our approach.

Models	ECSSD			DUT-OMRON			PASCAL-S			HKU-IS			DUTS-TE			XPIE		
	MAE	F_β^w	F_β	MAE	F_β^w	F_β	MAE	F_β^w	F_β	MAE	F_β^w	F_β	MAE	F_β^w	F_β	MAE	F_β^w	F_β
IPS	0.048	0.868	0.902	0.068	0.679	0.723	0.084	0.774	0.823	0.043	0.839	0.873	0.050	0.753	0.779	0.052	0.801	0.845
IPS + BLS	0.046	0.877	0.914	0.060	0.699	0.752	0.080	0.791	0.839	0.042	0.845	0.878	0.047	0.762	0.809	0.050	0.812	0.859
BANet	**0.035**	**0.908**	**0.929**	**0.059**	**0.736**	**0.763**	**0.072**	**0.810**	**0.849**	**0.032**	**0.886**	**0.905**	**0.040**	**0.811**	**0.829**	**0.044**	**0.839**	**0.873**

Table 3. Comparisons of different contextual integration modules on six datasets. "w/o ISD" represents BANet without ISD, "r/w ASPP" means ISD is replaced with ASPP, and "r/w ASPP" means ISD is replaced with ASPP-M.

Models	ECSSD			DUT-OMRON			PASCAL-S			HKU-IS			DUTS-TE			XPIE		
	MAE	F_β^w	F_β	MAE	F_β^w	F_β	MAE	F_β^w	F_β	MAE	F_β^w	F_β	MAE	F_β^w	F_β	MAE	F_β^w	F_β
w/o ISD	0.046	0.876	0.913	0.064	0.701	0.745	0.086	0.772	0.827	0.039	0.858	0.890	0.049	0.766	0.797	0.052	0.806	0.852
r/w ASPP	0.040	0.889	0.912	0.060	0.713	0.740	0.077	0.790	0.832	0.036	0.868	0.887	0.045	0.780	0.793	0.471	0.821	0.855
r/w ASPP-M	0.039	0.891	0.917	0.060	0.711	0.742	0.078	0.789	0.834	0.036	0.870	0.891	0.044	0.786	0.800	0.048	0.822	0.857
BANet	**0.035**	**0.908**	**0.929**	**0.059**	**0.736**	**0.763**	**0.072**	**0.810**	**0.849**	**0.032**	**0.886**	**0.905**	**0.040**	**0.811**	**0.829**	**0.044**	**0.839**	**0.873**

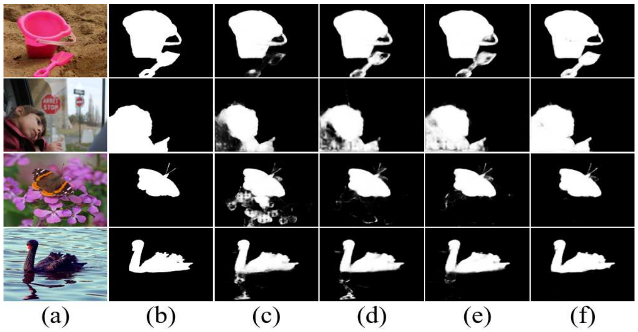


Figure 8. Comparisons of ASPP and our ISD based on the proposed boundary-ware network. (a) Images; (b) ground-truth; (c) without ISD; (d) ASPP as a replacement of ISD; (e) ASPP-M as a replacement of ISD; (f) Our approach.

IPS, BLS and TCS, the proposed method can generate more accurate results.

Effectiveness of Integrated Successive Dilation Module. Atrous Spatial Pyramid Pooling (ASPP) [4] is a common module for semantic segmentation [4, 5, 44], which consists of multiple parallel convolutional layers with filters at different dilation rates of [6, 12, 18, 24] , thus capturing feature receptive fields at different scales.

To validate the effectiveness of our ISD, we construct three different models based on our BANet to compare on six benchmark datasets. The first network is that we remove ISD from our BANet. Secondly, we replace ISD with ASPP in BANet as the second network. Moreover, for a fairer comparison, we modify ASPP denoted as "ASPP-M" with the same branches and same dilation rates like ISD except for the short information flows and replace ISD with ASPP-M in our BANet as the third network.

The comparison of the three models and our BANet is listed in Tab. 3. From Tab. 3, we find that after ISD is removed from BANet, the performance of the method decreases dramatically on all the six datasets due to the lack of the capability to extract features for a region embedded in various contexts. As shown in the third row of Fig. 8, the flowers close to the butterfly and the reflection of the goose in water are mistakenly detected as salient objects.

In fact, when the network utilizes ASPP or ASPP-M, the performance can also be improved to some extent. However, as the ISD has more information flow paths to aggregate the contextual information, the feature invariance can be enhanced better. Even if there are large appearance changes in the interiors of a large salient object, the whole salient object can be highlighted well, as shown in the last one line of Fig. 8.

5. Conclusion

In this paper, we revisit the problem of SOD from the perspective of selectivity-invariance dilemma, where feature selectivity and invariance are required by different regions in salient objects. To solve this problem, we propose a novel boundary-aware network with successive dilation for salient object detection. In this network, boundary localization and interior perception streams are introduced to capture features with selectivity and invariance, respectively. A transition compensation stream is adopted to amend the probable failures between boundaries and interiors. Then the output of these three streams are fused to obtain the saliency mask in a boundary-aware feature mosaic selection manner. Moreover, we also propose a novel integrated successive dilation module for enhancing feature invariance to help perceiving and localizing salient objects. Extensive experiments on six benchmark datasets have validated the effectiveness of the proposed approach.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under grant 2017YFB1002400, the National Natural Science Foundation of China (61672072, U1611461 and 61825101), and Beijing Nova Program under Grant Z181100006218063.

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *CVPR*, 2009. 2, 5

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

[3] Hao Chen and Youfu Li. Progressively complementarity-aware fusion network for rgb-d salient object detection. In *CVPR*, 2018. 3

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 40(4):834–848, 2018. 8

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 8

[6] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. Reverse attention for salient object detection. In *ECCV*, 2018. 3, 6

[7] Xiaowu Chen, Anlin Zheng, Jia Li, and Feng Lu. Look, perceive and segment: Finding the salient objects in images via two-stream fixation-semantic cnns. In *ICCV*, 2017. 1, 3, 5, 6

[8] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE TPAMI*, 37(3):569–582, 2015. 2, 5, 6

[9] Zijun Deng, Xiaowei Hu, Lei Zhu, Xuemiao Xu, Jing Qin, Guoqiang Han, and Pheng-Ann Heng. R3net: Recurrent residual refinement network for saliency detection. In *IJCAI*, 2018. 6

[10] Deng-Ping Fan, Ming-Ming Cheng, Jiang-Jiang Liu, Shang-Hua Gao, Qibin Hou, and Ali Borji. Salient objects in clutter: Bringing salient object detection to the foreground. In *ECCV*, 2018. 1

[11] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *CVPR*, 2019. 1

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 5, 6

[13] Seunghoon Hong, Tackgeun You, Suha Kwak, and Bohyung Han. Online tracking by learning discriminative saliency map with convolutional neural network. In *ICML*, 2015. 1

[14] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. *IEEE TPAMI*, 2019. 6

[15] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. Salient object detection: A discriminative regional feature integration approach. In *CVPR*, 2013. 1

[16] Dominik A Klein and Simone Frintrop. Center-surround divergence of feature statistics for salient object detection. In *ICCV*, 2011. 2

[17] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, 2011. 6

[18] Baisheng Lai and Xiaojin Gong. Saliency guided dictionary learning for weakly-supervised image parsing. In *CVPR*, 2016. 1

[19] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015. 1

[20] Gayoung Lee, Yu-Wing Tai, and Junmo Kim. Deep saliency with encoded low level distance map and high level features. In *CVPR*, 2016. 2, 6

[21] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *CVPR*, 2015. 1, 5, 7

[22] Guanbin Li and Yizhou Yu. Visual saliency detection based on multiscale deep cnn features. *IEEE TIP*, 25(11):5012–5024, 2016. 2, 6

[23] Xin Li, Fan Yang, Hong Cheng, Wei Liu, and Dinggang Shen. Contour knowledge transfer for salient object detection. In *ECCV*, 2018. 1, 3, 5, 6

[24] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *CVPR*, 2014. 1, 5

[25] Nian Liu and Junwei Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *CVPR*, 2016. 2

[26] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *CVPR*, 2018. 5, 6

[27] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *IEEE TPAMI*, 33(2):353–367, 2010. 6

[28] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015. 5

[29] Zhiming Luo, Akshaya Kumar Mishra, Andrew Achkar, Justin A Eichel, Shaozi Li, and Pierre-Marc Jodoin. Non-local deep features for salient object detection. In *CVPR*, 2017. 2, 6

[30] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *CVPR*, 2014. 5

[31] Zhixiang Ren, Shenghua Gao, Liang-Tien Chia, and Ivor Wai-Hung Tsang. Region-based saliency detection and its application in object recognition. *IEEE TCSVT*, 24(5):769–779, 2014. 1

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 5, 6

[33] Wei-Chih Tu, Shengfeng He, Qingxiong Yang, and Shao-Yi Chien. Real-time salient object detection with a minimum spanning tree. In *CVPR*, 2016. 2

[34] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, 2017. 1, 5, 6

[35] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Xiang Ruan. Salient object detection with recurrent fully convolutional networks. *IEEE TPAMI*, 2019. 6

[36] Tiantian Wang, Ali Borji, Lihe Zhang, Pingping Zhang, and Huchuan Lu. A stagewise refinement model for detecting salient objects in images. In *ICCV*, 2017. 1, 2, 5, 6

[37] Tiantian Wang, Lihe Zhang, Huchuan Lu, Chong Sun, and Jinqing Qi. Kernelized subspace ranking for saliency detection. In *ECCV*, 2016. 3, 6

[38] Tiantian Wang, Lihe Zhang, Shuo Wang, Huchuan Lu, Gang Yang, Xiang Ruan, and Ali Borji. Detect globally, refine locally: A novel approach to saliency detection. In *CVPR*, 2018. 3, 5, 6

[39] Changqun Xia, Jia Li, Xiaowu Chen, Anlin Zheng, and Yu Zhang. What is and what is not a salient object? learning salient object detector by ensembling linear exemplar regressors. In *CVPR*, 2017. 1, 5

[40] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017. 6

[41] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, 2015. 3

[42] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *CVPR*, 2013. 1, 5

[43] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013. 1, 5

[44] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *CVPR*, 2018. 8

[45] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *ICLR*, 2016. 3

[46] Yu Zeng, Huchuan Lu, Lihe Zhang, Mengyang Feng, and Ali Borji. Learning to promote saliency detectors. In *CVPR*, 2018. 3

[47] Lu Zhang, Ju Dai, Huchuan Lu, You He, and Gang Wang. A bi-directional message passing model for salient object detection. In *CVPR*, 2018. 2

[48] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *ICCV*, 2017. 1, 2, 5, 6

[49] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Baocai Yin. Learning uncertain convolutional features for accurate saliency detection. In *ICCV*, 2017. 2, 5, 6

[50] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. Progressive attention guided recurrent network for salient object detection. In *CVPR*, 2018. 1, 2, 5, 6