Towards High-Resolution Salient Object Detection

Yi Zeng¹, Pingping Zhang¹, Jianming Zhang², Zhe Lin², Huchuan Lu¹*
¹ Dalian University of Technology, China
² Adobe Research, USA
{dllgzy, jssxzhpp}@mail.dlut.edu.cn, {jianmzha, zlin}@adobe.com, lhchuan@dlut.edu.cn

Abstract

Deep neural network based methods have made a significant breakthrough in salient object detection. However, they are typically limited to input images with low resolutions (400 × 400 pixels or less). Little effort has been made to train neural networks to directly handle salient object segmentation in high-resolution images. This paper pushes forward high-resolution saliency detection, and contributes a new dataset, High-Resolution Salient Object Detection (HRSOD). To our best knowledge, HRSOD is the first high-resolution saliency detection dataset to date. As another contribution, we also propose a novel approach, which incorporates both global semantic information and local high-resolution details, to address this challenging task. More specifically, our approach consists of a Global Semantic Network (GSN), a Local Refinement Network (LRN) and a Global-Local Fusion Network (GLFN). GSN extracts the global semantic information based on downsampled entire image. Guided by the results of GSN, LRN focuses on some local regions and progressively produces high-resolution predictions. GLFN is further proposed to enforce spatial consistency and boost performance. Experiments illustrate that our method outperforms existing state-of-the-art methods on high-resolution saliency datasets by a large margin, and achieves comparable or even better performance than them on some widely used saliency benchmarks.

1. Introduction

Salient object detection, aiming at accurately detecting and segmenting the most distinctive object regions in a scene, has drawn increasing attention in recent years [8, 46, 47, 51, 48]. It is regarded as a very important task that can facilitate a wide range of applications, such as image understanding [20, 53, 44], object segmentation [18], image captioning [10, 7, 40] and light field 3D display [35].

*Corresponding author.
for high-resolution saliency detection task. Meanwhile, there is little research effort to train neural networks to directly handle salient object segmentation in high-resolution images.

However, this line of work is very important since it can inspire or enable many practical tasks such as image editing [31, 39, 23], medical image analysis [4], etc. Specifically, when served as a pre-processing step of background replacement and depth-of-field, high-resolution salient object detection should be as accurate as possible to provide users with realistic composite images [29]. If the predicted boundaries are not accurate, there may be artifacts which certainly affect users’ experience. Thus, this paper pushes forward the task of high-resolution salient object detection.

To our best knowledge, our approach is the first work for high-resolution salient object detection. Since there is no high-resolution training and test dataset for saliency detection, we contribute a new dataset, High-Resolution Salient Object Detection (HRSOD). More details about our HRSOD will be presented in Section 3.

As for developing high-resolution saliency detection methods, there are three intuitive methods. The first is simply increasing the input size to maintain a relative high resolution and object details after a series of pooling operations. However, the large input size results in significant increases in memory usage. Moreover, it remains a question that if we can effectively extract details from lower-level layers in such a deep network through back propagation. The second method is partitioning inputs into patches and making predictions patch-by-patch. However, this type of method is time-consuming and can easily be affected by background noise. The third one includes some post-processing methods such as CRF [19] or graph cuts [28], which can address this issue to a certain degree. But very few works attempted to solve it directly within the neural network training process. As a result, the problem of applying DNNs for high-resolution salient object detection is fairly unsolved.

To address above issues, we propose a novel deep learning approach for high-resolution salient object detection without any post-processing. It has a Global Semantic Network (GSN) for extracting semantic information and a Local Refinement Network (LRN) for optimizing local details. A global semantic guidance is introduced from GSN to LRN in order to ensure global consistency. Besides, an Attended Patch Sampling (APS) scheme is proposed to enforce LRN to focus on uncertain regions, and this scheme provides a good trade-off between performance and efficiency. Finally, a Global-Local Fusion Network (GLFN) is proposed to enforce spatial consistency and further boost performance at high resolution.

To summarize, our contributions are as follows:

- We contribute the first high-resolution salient object detection dataset (HRSOD) with rich boundary details and accurate pixel-wise annotations.
- We provide a new paradigm for high-resolution salient object detection which first uses GSN for extracting semantic information, and a guided LRN for optimizing local details, and finally GLFN for prediction fusion.
- We perform extensive experiments to demonstrate that our method outperforms other state-of-the-art methods on high-resolution saliency datasets by a large margin, and achieves comparable performance on some widely used saliency benchmarks.

2. Related Work

In the past few decades, lots of approaches have been proposed to solve the saliency detection problem. Early researches are mainly based on low-level features, such as image contrast [16, 6], texture [41, 42] and background prior [22, 37]. These models are efficient and effective in simple scenarios, but they are not always robust in handling challenging cases. A detailed survey of these methods can be found in [2].

More recently, learning based saliency detection methods have achieved expressive performance, and they can coarsely be divided into two categories, i.e., patch-based saliency and FCN-based saliency.

2.1. Patch-based Saliency

Existing patch-based methods make saliency prediction for each image patch. For example, Wang et al. [32] present a saliency detection algorithm by integrating both local estimation and global search. Then, Li et al. [21] propose to utilize multi-scale features in multiple generic CNNs to predict the saliency degree of each superpixel. With the same purpose of predicting the saliency degree of each superpixel, Zhao et al. [52] use a multi-context deep CNN to predict saliency maps taking global and local context into account. The above methods include several fully connected layers to make predictions in superpixel-level, resulting in expensive computational cost and the loss of spatial information. What’s more, all of them make very coarse predictions and lack low-level details.

2.2. FCN-based Saliency

Liu et al. [24] design a deep hierarchical saliency network and progressively recover image details via integrating local context information. Zhang et al. [49] propose a generic framework to integrate multi-level features into different resolutions for finer saliency maps. In order to better integrate features from different levels, Zhang et al. [45] propose a bi-directional message passing module with a gate function to integrate multi-level features. Wang et al. [36] use a boundary refinement network to learn propagation coefficients for each spatial position.
Lots of research efforts have been made to recover image details in final predictions. However, for high-resolution images, all existing FCN-based methods down-sample the inputs, thus lose high-resolution details and fail to predict fine-grained saliency maps.

Several researchers attempt to remedy this problem by using post-processing techniques for finer predictions. However, traditional CRF [19] and guided filtering are very time-consuming and their improvement is very limited. Wu et al. [38] propose a more efficient guided filtering layer. However, their performance is just comparable with the CRF. To reduce this gap, we propose a method to combine the advantages of patch-based methods (maintaining details and saving memory) and FCN-based methods (having rich contextual information).

3. High-Resolution Saliency Detection Dataset

There exist several datasets for saliency detection, but none of them is specifically designed for high-resolution salient object detection. Three main drawbacks are apparent. First, all images in current datasets have extremely limited resolutions. Concretely, the longest edge of each image is less than 500 pixels. These low-resolution images are not representative for today’s image processing applications. Second, to relieve the burden of users, it is essential to output masks with extremely high accuracy in boundary regions. But images in existing saliency detection datasets are inadequate in providing rich object boundary details for training DNNs. In addition, widely used saliency datasets also have some problems in annotation quality, such as failing to cover all saliency regions (Figure 2 (c)), including background disturbance into foreground annotation (Figure 2 (d)), or low contour accuracy (Figure 2 (e)).

To address the above urgent issues, we contribute a High-Resolution Salient Object Detection (HRSOD) dataset, containing 1610 training images and 400 test images. The total 2010 images are collected from the website of Flickr\(^1\) with the license of all creative commons. Pixel-level ground truths are manually annotated by 40 subjects. The shortest edge of each image in our HRSOD is more than 1200 pixels. Figure 2 presents the image size comparison between our HRSOD and existing saliency detection datasets. For existing datasets, we only show the results on HKU-IS dataset [21] (The maximum is less than 600.). (b) The histogram of diagonal length on our HRSOD (The minimum is over 1000.). (c)-(f) Sample images from various dataset, with ground truth masks overlaid. Concretely, (c) is from HKU-IS [21]. (d) is from DUTS-Test [33]. (e) is from THUR [5]. And (f) is an example of our HRSOD. Best viewed by zooming in.

Aiming at high-resolution salient object detection, the whole dataset is publicly available\(^2\).

4. Our Method

In this paper, we propose a novel method for detecting salient objects in high-resolution images with limited GPU memory. Our framework includes three branches, i.e., Global Semantic Network (GSN), Local Refinement Network (LRN) and Global-Local Fusion Network (GLFN). Figure 3 shows an overall illustration of the proposed approach. GSN aims at extracting semantic knowledge in a global view. Guided by GSN, LRN is designed to refine uncertain sub-regions. Finally, GLFN takes high-resolution images as inputs and further enforces spatial consistency of the fused predictions from GSN and LRN.

To be specific, let \( \{ X_i = (I_i, L_i) \}_{i=1}^N \) be the training set, containing both the training image \( I_i \) and its pixel-wise saliency label \( L_i \). The input image \( I_i \) is first fed forward through GSN to obtain a coarse saliency map \( F_i \), denoted

---

\(^1\)https://www.flickr.com

\(^2\)https://github.com/yi94code/HRSOD
where $DS(\cdot)$ denotes down-sampling images to $384 \times 384$ while $UP(\cdot)$ denotes up-sampling predictions to original size. $\theta$ denotes all parameters in GSN. Then $I_i$ is put into our proposed Attended Patch Sampling (APS) scheme (Algorithm 1) to generate sub-images $\{P_{m,I}^i\}_{m=1}^M$, which are attended to uncertain regions ($M$ is the total number of sub-images for each input $I_i$). Subsequently, each $P_{m,I}^i$ is fed forward through LRN to get a refined saliency map $R_{m,I}^F$. Semantic guidance is introduced from GSN to LRN (Section 4.2). Finally, the outputs of GSN and LRN are fused and fed forward through GLFN for final prediction $S_i$. These two stages can be formulated as:

$$F_i = UP(GSN(DS(I_i), \theta))$$  \hspace{1cm} (1)

$$\{R_{m,I}^F\}_{m=1}^M = LRN(\{P_{m,I}^I\}_{m=1}^M, \phi)$$  \hspace{1cm} (2)

$$S_i = GLFN(I_i, \text{Fuse}(\{P_{m,I}^I\}_{m=1}^M, F_i, \psi))$$  \hspace{1cm} (3)

where $\phi$ and $\psi$ denote the parameters of LRN and GLFN, respectively. $\text{Fuse}(\cdot)$ denotes fusion operation (more details can be seen in Section 4.4).

### 4.1. Network Architecture for GSN and LRN

![Network Architecture](Image)

Figure 4. (a) Network architecture for both GSN and LRN. (b) Incorporate global guidance only for LRN.

We adopt the same backbone for GSN and LRN. Our model is simply built on the FCN architecture with the pre-trained 16-layer VGG network [30]. The original VGG-16 network [30] is trained for image classification task while our model is trained for saliency detection, a pixel-wise prediction task. Therefore, we simply abandon all layers after conv5-3 to maintain a higher resolution.

In order to enlarge receptive field, we employ dilated convolutional layers [43] to capture contextual information. Dilated convolution, also known as atrous convolution, has a superior ability to enlarge the field of view without increasing the number of parameters. As shown in Figure 4 (a), we add four dilated convolutional layers on the top of conv3-3, conv4-3 and conv5-3 in our revised VGG-16. All the dilated convolutional layers have the same kernel size and output channels, $i.e.$, $k = 3$ and $c = 32$. The rates of the four dilated convolutional layers in the same block are set with dilation $= 1, 3, 5, 7$ respectively.

To improve the output resolution, we first generate three saliency score maps through the last three blocks. Secondly, we add three additional deconvolutional layers, the first two of which have $2 \times$ upsampling factors and the last of which has a $4 \times$ upsampling factor. Thirdly, inspired by [25], we build two skip connections from the saliency score maps generated by block 3 and block 4 to combine high-level features with meaningful semantic information and low-level features with large amount of details (See Figure 4 (a)). More details are provided in the supplementary material.

### 4.2. Semantic Guidance from GSN to LRN

The saliency maps generated by GSN are based on the full image and embedded with rich contextual information. Nevertheless, due to its small input size of $384 \times 384$, lots of low-level details are lost, especially when the original images have very high resolutions ($e.g.$, $1920 \times 1080$). That is to say, it barely learns to capture saliency properties at a coarse scale. As a result, GSN is competent in giving a rough saliency prediction but insufficient to precisely localize salient objects. In contrary, LRN takes sub-images as input, avoiding down-sampling which results in the loss of details. However, since sub-images are too local to indicate which area is more salient, LRN may be confused about which region should be highlighted. Also, LRN alone may have false alarms in some locally salient regions. Therefore, we propose to introduce the semantic guidance from GSN to LRN, in order to enhance global contextual knowledge while maintain high-resolution details.

Specifically, we incorporate global semantic guidance in the decoder part. As illustrated in Figure 4 (b), given the coarse result $F_i$ of GSN, a patch $P_{m,F}^i$ is first cropped according to the location of patch $P_{m,I}^I$ in LRN. Then we concatenate $P_{m,F}^i$ with the corresponding feature maps in LRN.

### 4.3. Focus on Uncertain Regions

Compared with previous patch-based methods, our LRN has a notable difference. Traditional patch-based methods usually infer every patch in the image by sliding window
or superpixels, which is extremely time-consuming. We note that GSN has already succeeded to assign most pixels with right labels. Therefore, LRN only needs to focus on harder regions. Such a hierarchical prediction manner (GSN for easy regions and LRN for harder regions) makes our method more efficient and accurate. An Attended Patch Sampling (APS) scheme is proposed for this task. Guided by the results of GSN, it can generate sub-images attended to uncertain regions. Algorithm 1 presents a rough procedure of APS (More details can be seen in supplementary material.). We use the attention map $A_i$ to indicate all uncertain pixels and it can be formulated as:

$$A_i(x, y) = \begin{cases} 1 & T_1 < F_i(x, y) < T_2 \\ 0 & otherwise \end{cases}$$  \hspace{1cm} (4)

In Algorithm 1, $w$ denotes the width of non-zero area in $A_i$, $X_L$ and $X_R$ are the $x$ coordinates of the leftmost and rightmost non-zero pixels in $A_i$. $n$ is a constant, which controls the overlapping between different patches. $r$ is a random numbers for generating sub-images with varied sizes. We have performed grid search for setting these hyper-parameters and found that the results were not sensitive to their specific choices. Therefore, we set them empirically in this work. We set $D = 384$, $n = 5$, $T_1 = 50$, $T_2 = 200$, and $r \in [-\frac{D}{6}, \frac{D}{6}]$ in all our experiments. Some image patches produced by APS are shown in Figure 6.

### 4.4. Global-Local Fusion Network

As illustrated in above sections, GSN and LRN are inherently complementary with each other. Our method leverages GSN to classify easy regions and LRN to refine harder ones. Then the final predictions can be obtained by fusing their results. A simple way to do this is to replace the saliency values of uncertain regions in $F_i$ (the result of GSN) by $\{R_m^l\}_{m=1}^M$ (the result of LRN). Overlapped areas will be averaged. However, this kind of fusion lacks spatial consistency and does not leverage rich details in original high-resolution images.

We propose to directly train a network to incorporate high-resolution information to help the fusion of GSN and LRN. To maintain all the high-resolution details from images, this network should not include any pooling layers or convolutional layers with large strides. With limited GPU memory, popular backbones (e.g., VGG and ResNet) can
Table 1. Quantitative comparisons with other state-of-the-arts in term of F-measure (larger is better) and MAE (smaller is better) on five datasets: THUR [5], HKU-IS [21] and DUTS [33]. THUR and HKU-IS are large-scale datasets, with 6232 and 4447 images, respectively. DUTS is a large saliency detection benchmark, contains 5019 test images.

<table>
<thead>
<tr>
<th>Method</th>
<th>HRSOD-Test</th>
<th>DAVIS-S</th>
<th>DUTS-Test</th>
<th>HKU-IS</th>
<th>THUR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$F_\beta$ &amp; S-m &amp; MAE &amp; $F_\beta$ &amp; S-m &amp; MAE &amp; $F_\beta$ &amp; S-m &amp; MAE &amp; $F_\beta$ &amp; S-m &amp; MAE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RFCN [34]</td>
<td>0.530</td>
<td>0.608</td>
<td>0.121</td>
<td>0.728</td>
<td>0.842</td>
</tr>
<tr>
<td>DHS [24]</td>
<td>0.746</td>
<td>0.848</td>
<td>0.059</td>
<td>0.774</td>
<td>0.865</td>
</tr>
<tr>
<td>UCF [50]</td>
<td>0.700</td>
<td>0.819</td>
<td>0.095</td>
<td>0.648</td>
<td>0.827</td>
</tr>
<tr>
<td>Amulet [49]</td>
<td>0.717</td>
<td>0.830</td>
<td>0.075</td>
<td>0.755</td>
<td>0.848</td>
</tr>
<tr>
<td>NLDF [26]</td>
<td>0.763</td>
<td>0.853</td>
<td>0.055</td>
<td>0.718</td>
<td>0.858</td>
</tr>
<tr>
<td>DSS [14]</td>
<td>0.756</td>
<td>0.840</td>
<td>0.060</td>
<td>0.728</td>
<td>0.865</td>
</tr>
<tr>
<td>RAS [3]</td>
<td>0.773</td>
<td>0.842</td>
<td>0.058</td>
<td>0.763</td>
<td>0.867</td>
</tr>
<tr>
<td>DGRL [36]</td>
<td>0.789</td>
<td>0.848</td>
<td>0.053</td>
<td>0.772</td>
<td>0.859</td>
</tr>
<tr>
<td>DGF [38]</td>
<td>0.795</td>
<td>0.824</td>
<td>0.058</td>
<td>0.785</td>
<td>0.847</td>
</tr>
<tr>
<td>Ours-D</td>
<td>0.857</td>
<td>0.876</td>
<td>0.040</td>
<td>0.850</td>
<td>0.875</td>
</tr>
<tr>
<td>Ours-DH</td>
<td>0.888</td>
<td>0.897</td>
<td>0.030</td>
<td>0.888</td>
<td>0.876</td>
</tr>
</tbody>
</table>

5. Experimental
5.1. Experimental Setup
5.1.1 Datasets

High-Resolution Saliency Detection Datasets. We mainly use our proposed HRSOD-Test to evaluate the performance of our method along with other state-of-the-art methods. To enrich the diversity, we also collect 92 images which are suitable for saliency detection from DAVIS [27], a densely annotated high-resolution video segmentation dataset. Images in this dataset are precisely annotated and have very high resolutions (i.e., 1920 × 1080). We ignore the categories of the objects and generate saliency ground truth masks for this dataset. For convenience, the collected dataset is denoted as DAVIS-S.

Low-Resolution Saliency Detection Datasets. In addition, we evaluate our method on three widely used benchmark datasets: THUR [5], HKU-IS [21] and DUTS [33]. THUR and HKU-IS are large-scale datasets, with 6232 and 4447 images, respectively. DUTS is a large saliency detection benchmark, contains 5019 test images.

5.1.2 Evaluation Metrics

We use four metrics to evaluate all methods: Precision-Recall (PR) curves, $F_\beta$ measure, Mean Absolute Error (MAE) and structure-measure [9]. PR curves are generated by binarizing the saliency map with a varied threshold from 0 to 255, then comparing the binary maps with the ground truth. $F_\beta$ measure is defined as $F_\beta = \frac{(1+\beta^2) \cdot \text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$. The precision and recall are computed under the threshold of twice the mean saliency value. $\beta$ is set to 0.3 as suggested in [1] to emphasize precision. MAE measures the average error of saliency maps. Structure-measure simultaneously evaluates region-aware and object-aware structural similarity between a saliency map and a ground truth mask. For detailed implementations, we refer readers to [9].

5.1.3 Implementation Details

All experiments are conducted on a PC with an i7-8700 CPU and a 1080 Ti GPU, with the Caffe toolbox [17]. In our method, every stage is trained to minimize a pixelwise softmax loss function, by using the stochastic gradient descent (SGD). Empirically, the momentum parameter is set to 0.9 and the weight decay is set to 0.0005. For GSN and LRN, the inputs are first warped into $3\times3 \times 384$ pixels. Therefore, we propose a lightweight network, name as Global-Local Fusion Network (GLFN). As shown in Figure 5, high-resolution RGB images and combined maps from GSN and LRN are concatenated together to be the inputs of GLFN. GLFN consists of some convolution layers with dense connectivity as in [15]. We set the growth rate $g$ to be 2 for saving memory. Similar to [15], we let the bottleneck layers ($1 \times 1$ convolution) produce $4g$ feature maps. On the top of these densely connected layers, we add four dilated convolutional layers to enlarge receptive field. All the dilated convolutional layers have the same kernel size and output channels, i.e., $k = 3$ and $c = 2$. The rates of the four dilated convolutional layers are set with $dilation = 1, 6, 12, 18$ respectively. At last, a $3 \times 3$ convolution is appended for final prediction. What is worth mentioning is that our proposed GLFN has an extremely small model size (i.e., 11.9 kB).
5.2. Comparison with the State-of-the-arts

We compare our algorithm with 9 state-of-the-art methods, including RFCN [34], DHS [24], UCF [50], Amulet [49], NLDF [26], DSS [14], RAS [3], DGF [38] and DGRL [36]. For a fair comparison, we use either the implementations with recommended parameter settings or the saliency maps provided by the authors. To demonstrate the effectiveness of our approach, we provide two versions of our results. Ours-D represents for training on DUTS while Ours-DH represents for training on DUTS and HRSOD.

One thing deserves to be mentioned is that in our framework, GSN and LRN can be any saliency detection model. We just choose simple FCNs to validate the effectiveness of our framework. With our method, even simple FCNs can outperform other complicated models.

Quantitative Evaluation. $F_\beta$ measure, S-measure and MAE scores are given in Table 1. As can be seen, our method outperforms all the existing state-of-the-art methods on our new-built high-resolution datasets with a large margin. It also achieves comparable or even superior performance than them on some widely used saliency detection datasets. We provide the PR curves in the supplementary material due to limited space.

Qualitative Evaluation. Figure 8 shows a visual comparison of our method with respect to others. It can be seen that our method is capable of accurately detecting salient objects as well as suppressing the background clutter. Further, our saliency maps have better boundary shape and are much closer to the ground truth maps in various challenging scenarios.

5.3. Ablation Analysis and Discussion

5.3.1 Ablation Analysis

In this section, we provide the results about different variants of our method to further verify our main contributions.

LRN, GLFN vs CRF. In our method, LRN learns to refine uncertain regions under the guidance of GSN. To demonstrate its effectiveness, We also compare it with CRF [19], a widely used post-processing for saliency detection. The parameters are set as in [14]. We employ the CRF to refine predictions of GSN and LRN, denoted as GSN+CRF and GSN+APS+LRN+CRF, respectively. The results in Table 4 show that our method outperforms CRF by a large margin. Figure 8 shows the qualitative results. We find that our LRN and GLFN progressively improve details of saliency maps while the CRF fails to recover lost details.

APS vs RPS. To demonstrate the effectiveness of the proposed APS scheme, we train LRN on patches which are randomly sampled. For fair comparison, we set the number and size of sampled patches to be the same with our proposed APS. We denote this setting as GSN+RPS+LRN.
Table 2. The Boundary Displacement Error (smaller is better) of the state-of-the-art methods on high-resolution datasets. The best results are shown in bold.

<table>
<thead>
<tr>
<th>Network Structure</th>
<th>Ours*</th>
<th>Ours</th>
<th>DGF</th>
<th>DGRl</th>
<th>RPS</th>
<th>DSS</th>
<th>RLDF</th>
<th>RFCN</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSN+APS+LRN+GLFN</td>
<td>0.888</td>
<td>0.897</td>
<td>0.030</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Various metrics in Table 4 demonstrate that APS significantly outperforms RPS, which indicates the important role of our proposed APS.

**Performance vs number of patches.** For traditional patch-based methods, refining more patches brings more performance gain, but results in more computational cost. It seems like a tricky trade-off problem. Our proposed APS can ingeniously relieve this problem thanks to its focusing on uncertain regions. Figure 9 shows that our APS is less sensitive to number of patches.

![Figure 9. Refinement quality versus patch of numbers for different approaches. (a) S-measure versus patch of numbers. (b) MAE versus patch of numbers. Results are measured on the outputs of LRN.](image)

### 5.3.2 More Discussion

**Running time and model size.** Table 3 shows a comparison of running time and model size. Since other methods can not directly handle high-resolution images, the running time analysis of the compared methods is conducted with the same input size (384 × 384) for fair. Also, we provide our running time for 1024 × 1024 inputs, denoted as Ours*.

As it can be seen, our method is the fastest among all the compared methods and is quite efficient when directly handling high-resolution images.

**Boundary quality.** To further evaluate the precision of boundaries, we compare different methods by the Boundary Displacement Error (BDE) metric [11]. This metric measures the average displacement error of boundary pixels between two predictions, which can be formulated as:

\[
BDE(X, Y) = \frac{1}{2} \left[ \frac{1}{N_X} \sum_{x} \inf_{y \in Y} d(x, y) + \frac{1}{N_Y} \sum_{y} \inf_{x \in X} d(x, y) \right]
\]

where \(X\) and \(Y\) are two boundary pixel sets, and \(x, y\) are pixels in them, respectively. \(N_X\) and \(N_Y\) denote the number of pixels in \(X\) and \(Y\). \(\inf\) represents for the infimum and \(d(\cdot)\) denotes Euclidean distance. We only compute the BDE on high-resolution datasets because other benchmarks are not qualified enough on boundaries in pixel-level due to relatively poor annotations. The BDE for the state-of-the-art methods on HRSOD-Test and DAVIS-S are listed in Table 2. The results indicate that our predictions have better boundary shape and are much closer to the ground truth maps.

### 6. Conclusion

In this paper, we push forward high-resolution saliency detection task and provide a high-resolution saliency detection dataset (HRSOD) for facilitating studies in high-resolution saliency prediction. A novel approach is proposed to address this challenging task. It leverages both global semantic information and local high-resolution details to accurately detect salient objects in high-resolution images. Extensive evaluations on high-resolution datasets and popular benchmark datasets verify the effectiveness of our method. We will explore to develop weakly supervised high-resolution salient object detection in the future.

**Acknowledgements.** This paper is supported in part by National Natural Science Foundation of China No. 61725202, 61829102, 61751212, in part by the Fundamental Research Funds for the Central Universities under Grant No. DUT19GJ201 and gifts from Adobe.
References

[27] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and