

Towards High-Resolution Salient Object Detection

Yi Zeng¹, Pingping Zhang¹, Jianming Zhang², Zhe Lin², Huchuan Lu^{1*}

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 deep neural networks to directly handle salient object detection in very high-resolution images. This paper pushes forward high-resolution saliency detection, and contributes a new dataset, named 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 down-sampled 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 widely-used saliency benchmarks. The HRSOD dataset is available at <https://github.com/yi94code/HRSOD>.

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, 47, 46, 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

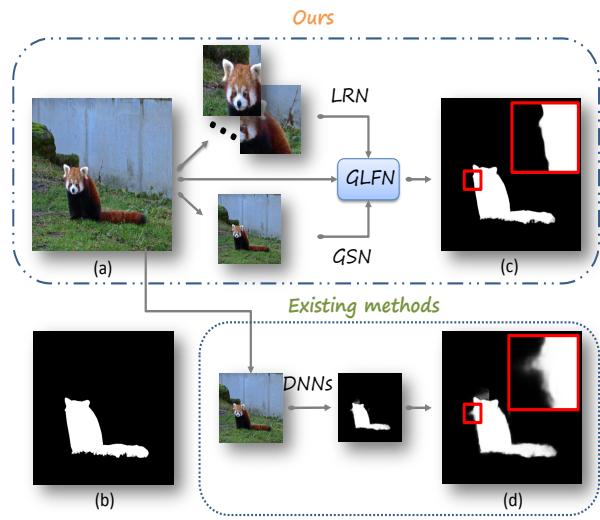


Figure 1. Pipeline comparison with state-of-the-art methods. (a) Input image. (b) Ground truth mask. (c) Our method. (d) Amulet [49]. Best viewed by zooming in.

captioning [10, 7, 40] and light field 3D display [35].

Deep Neural Networks (DNNs), e.g., VGG [30], ResNet [13], have achieved remarkable success in computer vision tasks using the typical input size such as 224×224 , 384×384 , etc. For most applications, such as image classification, object detection and visual tracking, the typical input size is enough to obtain satisfied results. For dense prediction tasks, e.g., image segmentation and saliency detection, deep learning based approaches also show impressive performance. But the inherited defect is very apparent, i.e., blurry boundary. Many research efforts have been made to remedy this problem. For example, Zhang *et al.* [49] employ deep recursive supervision and integrate multi-level features for accurate boundary prediction. However, the improvement is not significant, as illustrated in Figure 1 (d).

Furthermore, the resolution of the images taken by electronic products (e.g., smartphones) becomes very high, e.g., 720p, 1080p and 4K. When processing high-resolution images, the above defect becomes more severe. The state-of-the-art saliency detection methods generally down-scale the inputs to extract semantic information. In this pro-

*Corresponding author.

cess, many details are inevitably lost. Thus, they are not suitable for high-resolution saliency detection task. Meanwhile, there is little research effort to train neural networks to directly handle salient object segmentation in very 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 global semantic information, and a Local Refinement Network (LRN) for optimizing local object 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 introduce the first high-resolution salient object de-

tection 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 **H**igh-**R**esolution **S**alient **O**bject **D**etection (HRSOD) dataset, containing 1610 training images and 400 test images. The total 2010 images are collected from the website of Flickr¹ 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], and the results hold the same on other datasets. Besides, we provide an analysis of shape complexity in supplementary material. Compared with existing saliency datasets, our HRSOD avoids low-level mistakes via careful check by over 5 subjects (an example shown in Figure 2 (f)). To our best knowledge, HRSOD is cur-

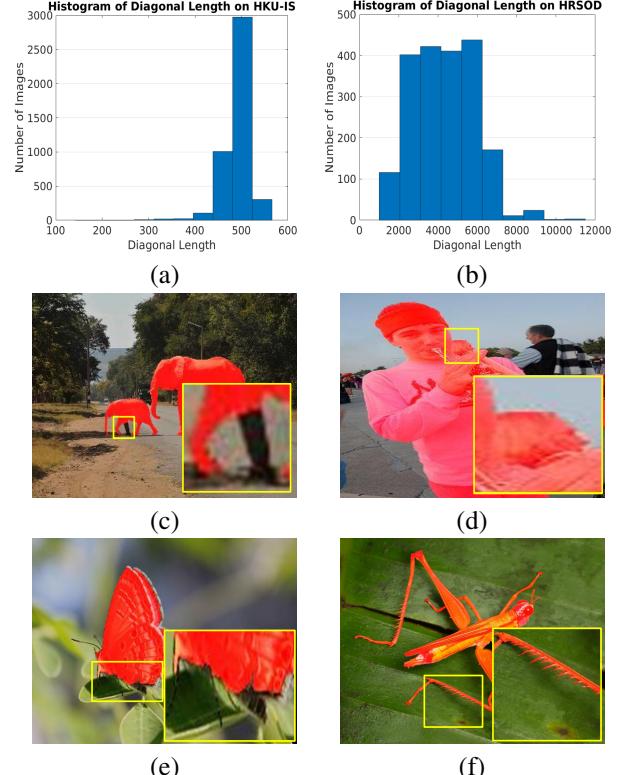


Figure 2. (a) The histogram of diagonal length on HKU-IS [21] (The maximum is less than 600.). (b) The histogram of diagonal length on our HRSOD (The minimum is over 1000.). (c)-(f) Sample images from various dataset, with ground truth masks overlaid. Concretely, (c) is from HKU-IS [21]. (d) is from DUTS-Test [33]. (e) is from THUR [5]. And (f) is an example of our HRSOD. Best viewed by zooming in.

rently the first high-resolution dataset for salient object detection. It is specifically designed for training and evaluating DNNs aiming at high-resolution salient object detection. The whole dataset is publicly available².

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 $\{\mathbf{X}_i = (\mathbf{I}_i, \mathbf{L}_i)\}_{i=1}^N$ be the training set, containing both the training image \mathbf{I}_i and its pixel-wise saliency label \mathbf{L}_i . The input image \mathbf{I}_i is first fed forward

¹<https://www.flickr.com>

²<https://github.com/yi94code/HRSOD>

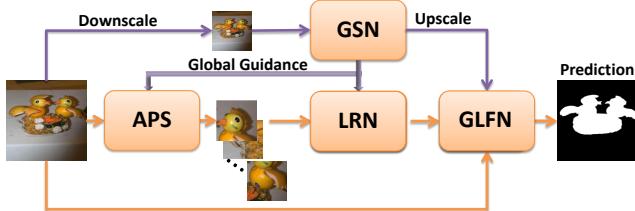


Figure 3. Overview of the network architecture. GSN and LRN takes downsampled entire images and attended sub-images as input, respectively. The guidance from GSN provides some semantic knowledge and ensures that our APS and LRN are attended to uncertain regions. A GLFN is appended to directly leverage high-resolution information to fuse the predictions from GSN and LRN.

through GSN to obtain a coarse saliency map \mathbf{F}_i , denoted as:

$$\mathbf{F}_i = UP(GSN(DS(\mathbf{I}_i), \theta)) \quad (1)$$

where $DS(\cdot)$ denotes down-sampling images to 384×384 while $UP(\cdot)$ denotes up-sampling predictions to original size. θ denotes all parameters in GSN. Then \mathbf{I}_i is put into our proposed Attended Patch Sampling (APS) scheme (Algorithm 1) to generate sub-images $\{\mathbf{P}_m^{\mathbf{I}_i}\}_{m=1}^M$, which are attended to uncertain regions (M is the total number of sub-images for each input \mathbf{I}_i). Subsequently, each $\mathbf{P}_m^{\mathbf{I}_i}$ is fed forward through LRN to get a refined saliency map $\mathbf{R}_m^{\mathbf{I}_i}$. 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:

$$\{\mathbf{R}_m^{\mathbf{I}_i}\}_{m=1}^M = LRN(\{\mathbf{P}_m^{\mathbf{I}_i}\}_{m=1}^M, \phi) \quad (2)$$

$$S_i = GLFN(\mathbf{I}_i, Fuse(\{\mathbf{R}_m^{\mathbf{I}_i}\}_{m=1}^M, \mathbf{F}_i), \psi) \quad (3)$$

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

4.1. Network Architecture for GSN and LRN

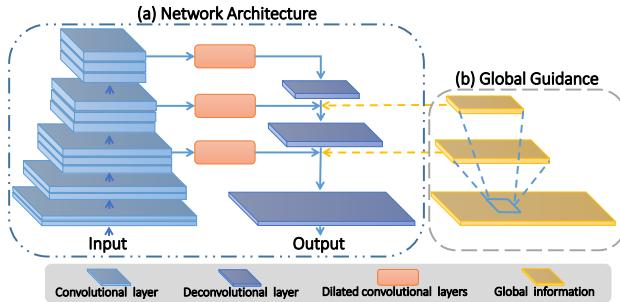


Figure 4. (a) Network architecture for both GSN and LRN. (b) Incorporate global guidance only for LRN.

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 384×384 , lots of low-level details are lost, especially when the original images have very high resolutions (*e.g.*, 1920×1080). 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 \mathbf{F}_i of GSN, a patch $\mathbf{P}_m^{\mathbf{F}_i}$ is first cropped according to the location of patch $\mathbf{P}_m^{\mathbf{I}_i}$ in LRN. Then we concatenate $\mathbf{P}_m^{\mathbf{F}_i}$ with the corresponding feature maps in LRN.

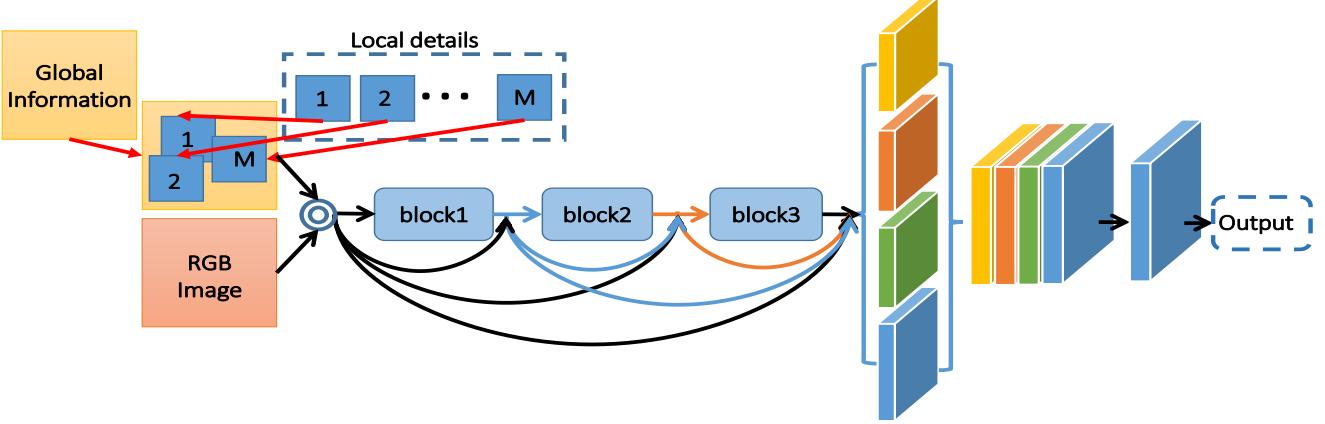


Figure 5. Global-Local Fusion Network.



Figure 6. Some sub-images produced by APS algorithm. (a) Original input image. (b)-(f) Typical sub-images produced by APS.

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:

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

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

Algorithm 1 Attended Patch Sampling.

Require: RGB image I_i , ground truth label L_i , base cropping size D .
Ensure: RGB patch set $\{P_m^{I_i}\}_{m=1}^M$, ground truth patch set $\{P_m^{L_i}\}_{m=1}^M$.

- 1: Generate attention map A_i from F_i , as in Equ. 4.
- 2: $N_x = \lceil w/D \rceil + n$
- 3: **for** $t = 1, \dots, N_x + 1$ **do**
- 4: $C = D + r$
- 5: $X_t = \min\{X_L + (t - 1) \times \lceil w/N_x \rceil, X_R\}$
- 6: $Y = \{y \mid A_i(X_t, y) = 1\}$
- 7: Pick out J pixels $(X_t, y(j))_{j=1}^J$ from (X_t, Y) .
- 8: Taking C as cropping size, $(X_t, y(j))_{j=1}^J$ as center pixels, crop $\{P_j^{I_i}\}_{j=1}^J$ and $\{P_j^{L_i}\}_{j=1}^J$ from I_i and L_i , respectively.
- 9: **end for**

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 $\{R_m^{I_i}\}_{m=1}^M$ (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

Method	HRSOD-Test			DAVIS-S			DUTS-Test			HKU-IS			THUR		
	F_β	S-m	MAE												
RFCN [34]	0.530	0.608	0.121	0.728	0.842	0.062	0.712	0.792	0.091	0.835	0.746	0.079	0.627	0.793	0.100
DHS [24]	0.746	0.848	0.059	0.774	0.865	0.034	0.724	0.817	0.067	0.855	0.746	0.053	0.673	0.803	0.082
UCF [50]	0.700	0.819	0.095	0.648	0.827	0.080	0.629	0.778	0.117	0.808	0.747	0.074	0.645	0.785	0.112
Amulet [49]	0.717	0.830	0.075	0.755	0.848	0.042	0.676	0.803	0.085	0.839	0.772	0.052	0.670	0.797	0.094
NLDF [26]	0.763	0.853	0.055	0.718	0.858	0.042	0.743	0.815	0.066	0.874	0.770	0.048	0.697	0.801	0.080
DSS [14]	0.756	0.840	0.060	0.728	0.865	0.041	0.791	0.822	0.057	0.895	0.779	0.041	0.731	0.801	0.073
RAS[3]	0.773	0.842	0.058	0.763	0.867	0.038	0.755	0.839	0.060	0.871	0.887	0.045	0.696	0.787	0.082
DGRL [36]	0.789	0.848	0.053	0.772	0.859	0.038	0.768	0.841	0.051	0.882	0.802	0.037	0.716	0.816	0.077
DGF [38]	0.795	0.824	0.058	0.785	0.847	0.037	0.776	0.803	0.062	0.893	0.869	0.043	0.734	0.799	0.070
Ours-D	0.857	0.876	0.040	0.850	0.875	0.029	0.796	0.827	0.052	0.891	0.882	0.042	0.740	0.820	0.067
Ours-DH	0.888	0.897	0.030	0.888	0.876	0.026	0.791	0.822	0.051	0.886	0.877	0.042	0.749	0.826	0.064

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 dataset. The best results are shown in bold.

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 not be trained with such a high-resolution input size (more than 1000×1000 pixels). Therefore, We propose a light-weighted 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 (1×1 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 3×3 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. Experiment

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 named 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_β 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_β 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. β^2 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 384×384 and the batch size is set to 32. The weights in block 1 to block 5 are initialized with the pre-trained VGG model [30], while

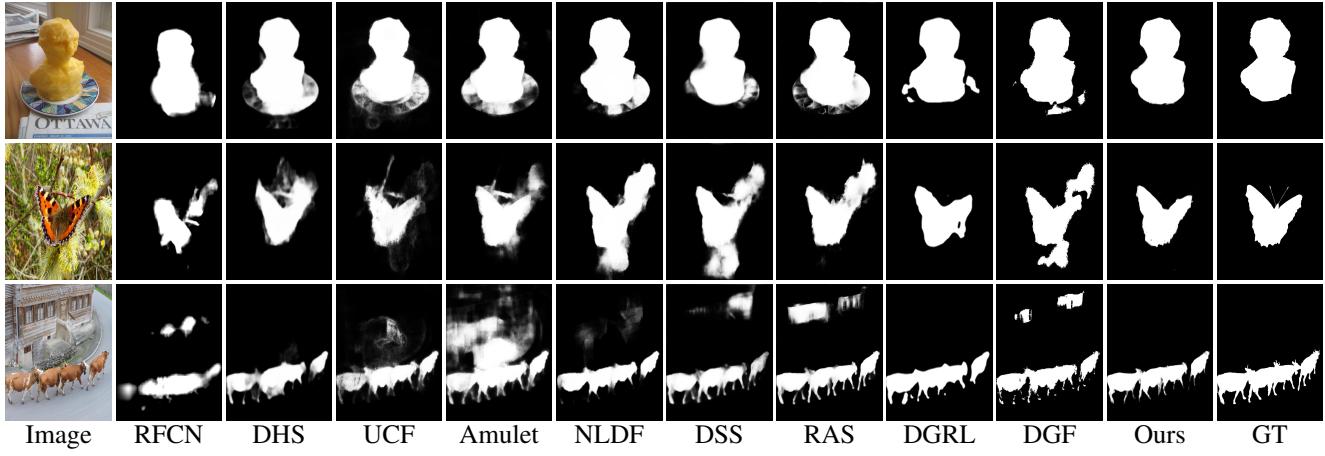


Figure 7. Visual comparison. All images are from HRSOD-Test dataset. Best viewed by zooming in.

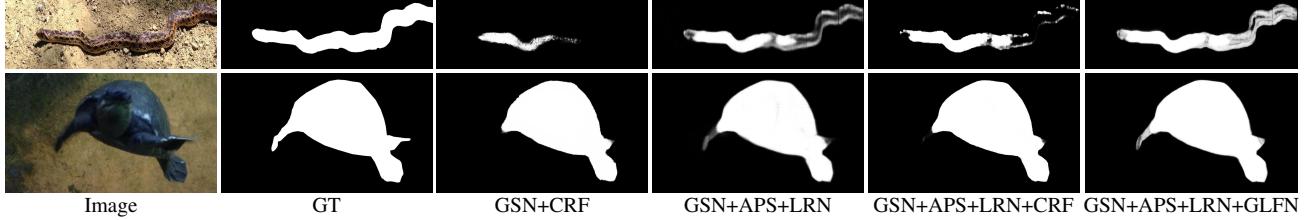


Figure 8. Visual comparison of our method with variations using Dense CRF [19].

weight parameters of newly-added convolutional layers are randomly initialized by using the “msra” method [12]. The learning rates of the pre-trained and newly-added layers are set to 1e-3 and 1e-2, respectively. GLFN is trained from scratch, and its weight parameters of convolutional layers are also randomly initialized by using the “msra” method. Its inputs are warped into 1024×1024 and the batch size is set to 2. Source code will be released.

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_β 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 per-

formance 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 challenging cases.

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 num-

Dataset	Ours	NLDF	UCF	DHS	DSS	Amulet	RAS	DGRL	DGF	RFCN
HRSOD-Test	17.57	22.34	22.84	22.85	25.53	25.75	26.26	30.10	32.91	68.98
DAVIS-S	8.18	23.56	15.69	18.34	19.35	21.11	18.49	14.48	19.77	21.00

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.

	Ours*	Ours	DGF	DGRL	RAS	DSS	NLDF	Amulet	UCF	DHS	RFCN
Time (s)	0.39	0.05	0.41	0.52	0.08	5.12	2.31	0.05	0.14	0.05	4.54
Model Size(MB)	129.6	129.6	248.9	648.0	81	447.3	425.9	132.6	117.9	376.2	1126.4

Table 3. Running time and model size of the state-of-the-art methods.

Network Structure	F_β	S-m	MAE
GSN	0.842	0.866	0.047
GSN+CRF	0.858	0.852	0.038
GSN+RPS+LRN	0.860	0.871	0.037
GSN+APS+LRN	0.877	0.883	0.036
GSN+APS+LRN+CRF	0.880	0.875	0.033
GSN+APS+LRN+GLFN	0.888	0.897	0.030

Table 4. Comparison of the different variants on HRSOD-Test.

ber and size of sampled patches to be the same with our proposed APS. We denote this setting as GSN+RPS+LRN. 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.

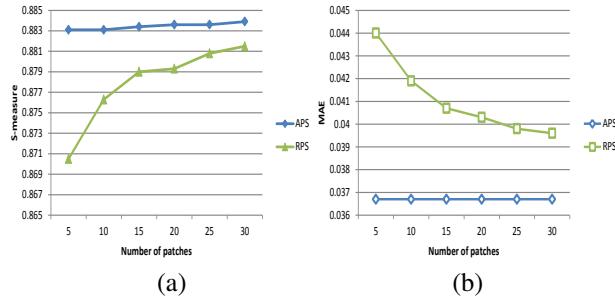


Figure 9. Refinement quality versus patch of numbers for different approaches. (a) S-measure vs. patch of numbers. (b) MAE versus patch of numbers. Results are measured on the outputs of LRN.

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:

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

where \mathbf{X} and \mathbf{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 \mathbf{X} and \mathbf{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 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

- [1] Ravi Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *CVPR*, pages 1597–1604, 2009. 6
- [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] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. Reverse attention for salient object detection. In *ECCV*, pages 234–250, 2018. 6, 7
- [4] Xu Chen, Bryan M Williams, Srinivasa R Vallabhaneni, Gabriela Czanner, Rachel Williams, and Yalin Zheng. Learning active contour models for medical image segmentation. In *CVPR*, pages 11632–11640, 2019. 2
- [5] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, and Shi-Min Hu. Salientshape: Group saliency in image collections. *The Visual Computer*, 30(4):443–453, 2014. 3, 6
- [6] 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
- [7] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *CVIU*, 163:90–100, 2017. 1
- [8] 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*, pages 186–202, 2018. 1
- [9] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A New Way to Evaluate Foreground Maps. In *ICCV*, pages 4548–4557, 2017. 6
- [10] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *CVPR*, pages 1473–1482, 2015. 1
- [11] Jordi Freixenet, Xavier Muñoz, David Raba, Joan Martí, and Xavier Cufí. Yet another survey on image segmentation: Region and boundary information integration. In *ECCV*, pages 408–422, 2002. 8
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *CVPR*, pages 1026–1034, 2015. 7
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1
- [14] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip Torr. Deeply supervised salient object detection with short connections. In *CVPR*, pages 5300–5309, 2017. 6, 7
- [15] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017. 6
- [16] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI*, 20(11):1254–1259, 1998. 2
- [17] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia*, pages 675–678, 2014. 6
- [18] Seong Joon Oh, Rodrigo Benenson, Anna Khoreva, Zeynep Akata, Mario Fritz, and Bernt Schiele. Exploiting saliency for object segmentation from image level labels. In *CVPR*, pages 4410–4419, 2017. 1
- [19] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, pages 109–117, 2011. 2, 3, 7
- [20] Baisheng Lai and Xiaojin Gong. Saliency guided dictionary learning for weakly-supervised image parsing. In *CVPR*, pages 3630–3639, 2016. 1
- [21] Guanbin Li and Yizhou Yu. Visual saliency based on multi-scale deep features. In *CVPR*, pages 5455–5463, 2015. 2, 3, 6
- [22] Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via dense and sparse reconstruction. In *ICCV*, pages 2976–2983, 2013. 2
- [23] Dani Lischinski, Zeev Farbman, Matt Uyttendaele, and Richard Szeliski. Interactive local adjustment of tonal values. In *ACM TOG*, volume 25, pages 646–653, 2006. 2
- [24] Nian Liu and Junwei Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *CVPR*, pages 678–686, 2016. 2, 6, 7
- [25] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 4
- [26] 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*, pages 6609–6617, 2017. 6, 7
- [27] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, pages 724–732, 2016. 6
- [28] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM TOG*, volume 23, pages 309–314, 2004. 2
- [29] Xiaoyong Shen, Aaron Hertzmann, Jiaya Jia, Sylvain Paris, Brian Price, Eli Shechtman, and Ian Sachs. Automatic portrait segmentation for image stylization. In *Computer Graphics Forum*, volume 35, pages 93–102, 2016. 2
- [30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 1, 4, 6
- [31] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, and Ming-Hsuan Yang. Sky is not the limit: semantic-aware sky replacement. *ACM TOG*, 35(4):149–162, 2016. 2
- [32] Lijun Wang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Deep networks for saliency detection via local estimation and global search. In *CVPR*, pages 3183–3192, 2015. 2

- [33] 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*, pages 136–145, 2017. [3](#), [6](#)
- [34] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Xiang Ruan. Saliency detection with recurrent fully convolutional networks. In *ECCV*, pages 825–841, 2016. [6](#), [7](#)
- [35] Shizheng Wang, Wenjuan Liao, Phil Surman, Zhigang Tu, Yuanjin Zheng, and Junsong Yuan. Salience guided depth calibration for perceptually optimized compressive light field 3d display. In *CVPR*, pages 2031–2040, 2018. [1](#)
- [36] 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*, pages 3127–3135, 2018. [3](#), [6](#), [7](#)
- [37] Yichen Wei, Fang Wen, Wangjiang Zhu, and Jian Sun. Geodesic saliency using background priors. In *ECCV*, pages 29–42, 2012. [2](#)
- [38] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. Fast end-to-end trainable guided filter. In *CVPR*, pages 1838–1847, 2018. [3](#), [6](#), [7](#)
- [39] Yunxuan Xiao, Yikai Li, Yuwei Wu, and Lizhen Zhu. Auto-retoucher (art)-a framework for background replacement and image editing. *arXiv preprint arXiv:1901.03954*, 2019. [2](#)
- [40] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015. [1](#)
- [41] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *CVPR*, pages 1155–1162, 2013. [2](#)
- [42] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, pages 3166–3173, 2013. [2](#)
- [43] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *ICLR*, 2015. [4](#)
- [44] Fan Zhang, Bo Du, and Liangpei Zhang. Saliency-guided unsupervised feature learning for scene classification. *IEEE TGRS*, 53(4):2175–2184, 2015. [1](#)
- [45] Lu Zhang, Ju Dai, Huchuan Lu, You He, and Gang Wang. A bi-directional message passing model for salient object detection. In *CVPR*, pages 1741–1750, 2018. [2](#)
- [46] Pingping Zhang, Wei Liu, Yinjie Lei, and Huchuan Lu. Hyperfusion-net: Hyper-densely reflective feature fusion for salient object detection. *PR*, 93:521–533, 2019. [1](#)
- [47] Pingping Zhang, Wei Liu, Huchuan Lu, and Chunhua Shen. Salient object detection with lossless feature reflection and weighted structural loss. *IEEE TIP*, 28(6):3048–3060, 2019. [1](#)
- [48] Pingping Zhang, Dong Wang, Huchuan Lu, and Hongyu Wang. Non-rigid object tracking via deep multi-scale spatial-temporal discriminative saliency maps. *arXiv:1802.07957*, 2018. [1](#)
- [49] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *ICCV*, pages 202–211, 2017. [1](#), [2](#), [6](#), [7](#)
- [50] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Baocai Yin. Learning uncertain convolutional features for accurate saliency detection. In *ICCV*, pages 212–221, 2017. [6](#), [7](#)
- [51] Pingping Zhang, Luyao Wang, Dong Wang, Huchuan Lu, and Chunhua Shen. Agile amulet: Real-time salient object detection with contextual attention. *arXiv:1802.06960*, 2018. [1](#)
- [52] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Saliency detection by multi-context deep learning. In *CVPR*, pages 1265–1274, 2015. [2](#)
- [53] Jun-Yan Zhu, Jiajun Wu, Yan Xu, Eric Chang, and Zhuowen Tu. Unsupervised object class discovery via saliency-guided multiple class learning. *IEEE TPAMI*, 37(4):862–875, 2015. [1](#)