Salient Object Detection with Pyramid Attention and Salient Edges

Wenguan Wang*[1], Shuyang Zhao *[2], Jianbing Shen†[1,2], Steven C. H. Hoi [3,4], Ali Borji [5]
[1]Inception Institute of Artificial Intelligence, UAE [2]Beijing Institute of Technology, China
[3]Singapore Management University, Singapore [4]Salesforce Research Asia, Singapore [5]MarkableAI, USA
wenguanwang.ai@gmail.com, zsyprich@bit.edu.cn
shenjianbingcg@gmail.com, chhoi@smu.edu.sg, aliborji@gmail.com

Abstract

This paper presents a new method for detecting salient objects in images using convolutional neural networks (CNNs). The proposed network, named PAGE-Net, makes two major novel contributions. The first is to devise an essential pyramid attention structure for salient object detection, which enables the network to concentrate more on salient regions while exploiting multi-scale saliency information. Such a stacked attention design offers a powerful way to efficiently enhance the representation ability of the corresponding network layer with an enlarged receptive field. The second contribution is to propose a salient edge detection module, which lies in the emphasis on the importance of salient edge information since it offers a strong cue to better segment salient objects and refine object boundaries. Such a salient edge detection module learns for precise salient boundary estimation, and thus encourages better edge-preserving salient object segmentation. Exhaustive experiments show that both of the proposed pyramid attention and salient edges are effective for salient object detection, and our PAGE-Net outperforms state-of-the-art approaches on several popular benchmarks with a fast inference speed (25FPS on a single GPU).

1. Introduction

Salient Object Detection (SOD) refers to the problem of locating and segmenting the most salient objects or regions in an image. It can be widely applied for improving a variety of vision tasks, such as object proposal generation [2], object segmentation [42, 44], photo cropping [45, 41], and video object tracking [13], among others. SOD has been extensively studied in computer vision. Traditional methods often design hand-crafted low-level features and make heuristic hypothesizes [49, 17], which often fail to yield satisfactory results for images with complex scenarios. Recently, deep learning approaches have emerged as an important trend for SOD and often reported significant improvements. Despite being studied actively, how to devise an effective yet efficient deep neural network model for SOD remains an open challenge.

In this paper, we propose a novel *Pyramid Attentive* and *salient edGE-aware* saliency model, named PAGE-Net, for saliency object detection, which is equipped with two key modules: (1) a *pyramid attention module* that efficiently enhances saliency representations by accounting for the multi-scale attention and enlarging receptive field of the saliency model; and (2) a *salient edge detection module* that explicitly learns salient object boundaries to better locate and sharpen salient objects. The design of the proposed PAGE-Net is motivated by the following two aspects.

First, feature representation is the crux of deep learning based saliency models, and it is always desirable to explore more efficient strategies for approaching scale-space feature learning problem. As witnessed in many saliency studies [34, 57, 14], multi-scale saliency features are crucial for SOD. As such, recent deep saliency models have mainly focused on combining the outputs from intermediate network layers. Unlike the existing work, we propose a novel pyramid attention model that inherits the feature-enhancing ability of attention mechanisms, and explicitly handles the problem of multi-scale saliency feature learning. Incorporating attention mechanisms into networks has proven useful for selecting task-relevant features [33]. As shown in Fig. 1, we extend attention mechanisms with hierarchical structures to enhance saliency computation. Such a design is significant because it efficiently increases *the receptive field* of the convolution layer (even for a shallow layer). Our saliency model is encouraged to focus on important regions using multi-scale information (Fig. 1 (b)). With pyramid attention, the background responses in the original features (Fig. 1 (c)) are successfully suppressed, leading to more discriminative saliency representations (Fig. 1 (d)) and better results (Fig. 1 (g)). Such an attention module also provides an additional dimension of interpretability by

*Equal contribution.
†Corresponding author: *Jianbing Shen*.

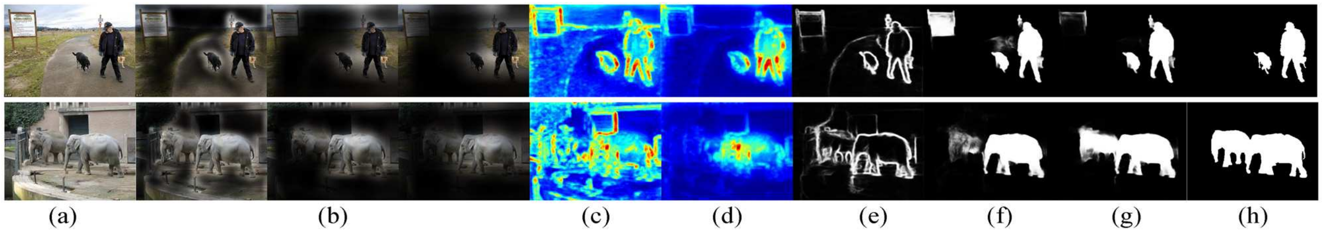


Figure 1: **Motivating examples and ideas for the proposed PAGE-Net.** (a) Image. (b) Pyramid attention maps (§3.1). (c) Original saliency features. (d) The refined saliency features via applying the proposed pyramid attention in (b). (e) Detected saliency edge map (§3.2). (f) Saliency results *w/o.* attention and salient edge detection modules. (g) Improved results via considering pyramid attention and salient edge cue. (h) Ground truth saliency map.

explaining where our saliency model is looking at.

Second, it is also desirable to find an effective means of enhancing the sharpness of salient object detection results. CNNs are designed to produce hierarchical feature maps through repeated pooling and subsampling operations, where higher layers gain larger receptive fields and stronger representation ability but loose much detailed spatial information. This can be useful for high-level tasks, but unfortunately degrades the accuracy of low-level tasks such as salient object segmentation where precise pixel-wise activations are required, especially on salient object boundaries. In the field of salient object detection, although densely connected or bottom-up/top-down network architectures [54, 14, 26] (see the scheme in Fig. 2 (a)) have been extensively studied to gradually recover salient object details in a top-down fashion, the issue of sharpness still remains a challenge. Inspired by recent advances in semantic segmentation [4, 6], we propose to equip saliency models with a salient-edge detection module, specially designed to detect the salient object boundaries. Thus, the network can leverage more explicit salient edges (Fig. 1 (e)) to better locate salient objects and sharpen the results (Fig. 1 (g)).

In summary, our main contributions are three-fold: **(i)** we present a pyramid attention model for discriminative saliency representations with multi-scale feature learning and an extended receptive field (§3.1); **(ii)** we propose a salient edge detection module that exploits salient edge information explicitly for salient object detection (§3.2); and **(iii)** we perform extensive experiments on six popular benchmarks, *i.e.*, ECCSD [49], DUT-OMRON [50], HKU-IS [21], PASCAL-S [25], SOD [30] and DUTS-TE [35], in which the proposed deep saliency model yields consistent improvements over a number of strong baselines. Finally, the proposed model runs very fast on modern GPUs, achieving a real-time inference speed of 25FPS.

2. Related Work

2.1. Salient Object Detection

The pioneering work for salient object detection can be dated back to Liu *et al.*, [28] and Achanta *et al.*, [1]. Since then, numerous subsequent works have been reported, mainly using *contrast based assumption* [9, 49, 17] and *background prior* [46, 58]. These early methods [43, 10] often heavily rely on hand-crafted features and heuristic hypothesizes.

Recently, due to the great successes of CNNs in computer vision, deep learning has emerged as a promising alternative for SOD. CNN-based saliency models allow flexible saliency representations with a powerful end-to-end learning ability, thus achieving significantly better performance than classic methods. A variety of deep learning approaches have been proposed in literature. For example, some methods integrate deep learning models with hand-crafted features [20], heuristic saliency priors [36], level set [15], contextual information [57], or explicit visual fixation [40]. Other methods leverage global and local saliency information [21, 34, 54, 29], combine pixel- and segment-level features [22], inspire connections between network layers [14], or explore more complex deep architectures [18, 26, 55, 37, 32].

One distinct difference of our method from the existing studies lies in the salient-edge-preservation property. Current saliency network architectures tend to stack multi-layer features. Although the final prediction layer accesses multi-scale and multi-level information and produces more precise saliency segmentation, the issue of sharpening remains unsolved due to the smoothness of convolution kernel and downsampling of spatial pooling. Some post-processing heuristics [36, 14, 22] have been adopted, but few explores how to embed salient edge information into a deep saliency model via end-to-end training. A few recent methods [53, 23] also explored the boundary cues, but they are very different from ours. For example, Zhang [53] *et al.* simply used an extra loss to emphasize the detection error for the pixels within the salient object boundaries. In [23], they considered semantic contour information from a pretrained contour detector [51]. By contrast, we extend each side-out layer with a salient edge detection module and learn the combination of edge and object information end-to-end.

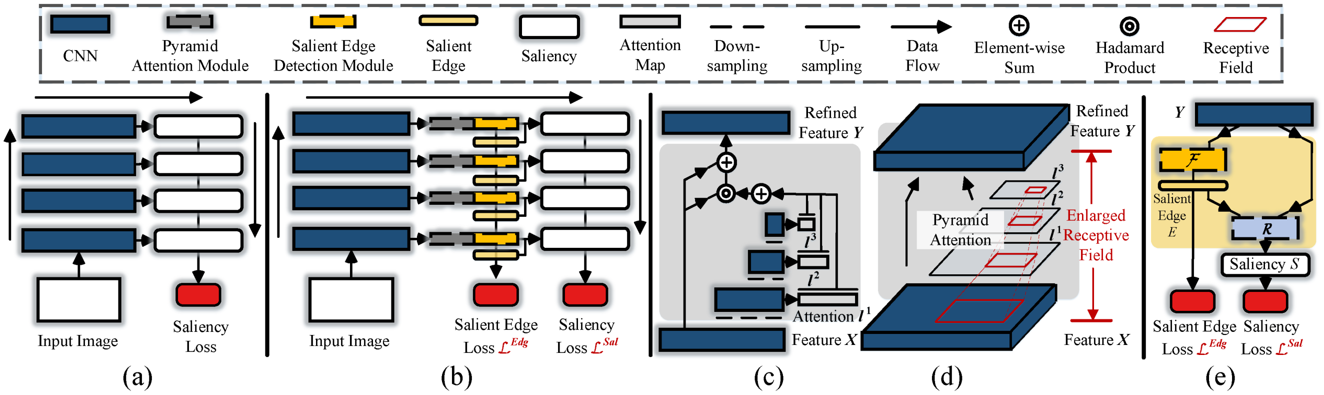


Figure 2: **Architecture designs of the proposed PAGE-Net.** (a) Typical bottom-up/top-down network architecture used in previous saliency methods. (b) PAGE-Net is equipped with two essential modules: pyramid attention module and salient edge detection module. (c) Architecture of the pyramid attention module (§3.1), where the attention is learned for enhancing saliency representation in multi-scales. (d) The pyramid attention module assigns corresponding convolution layer a global view with increased receptive field. (e) The edge detection module (§3.2) offers explicit edge information, which is used for locating salient objects and sharping salient object boundary.

2.2. Trainable Attention Mechanism in Network

Attention mechanisms of deep neural networks have been actively studied recently, which was first proposed by Bahdanau *et al.* [3] for neural machine translation. Later, it was proven useful in many natural language processing and vision tasks, *e.g.*, caption generation [48], question answering [52], and scene recognition [5, 33], among others. In such studies, attention is learned in an automatic, top-down, and goal-driven way, allowing the network to focus on the most task-relevant parts of images or sentences. Only a few very recent methods for SOD [56, 27, 7] employ attention networks. But our approach very differs from theirs, as they often only consider a single-layer attention design. In our approach, for each convolution layer, a pyramid of attentions is equipped for essentially learning to assign higher importance to salient regions while simultaneously addressing the issue of multi-scale learning. More importantly, such a pyramid attention design enables our model with a global view and improved learning ability via an enlarged receptive filed.

3. Our Method

Fig. 2 (b) gives a simplified illustration of PAGE-Net, which consists of three components: a backbone network for feature extraction, a pyramid attention module, and a salient edge detection module. We begin by describing our pyramid attention module (▭ in Fig. 2 (b)) in §3.1. A detailed description of our salient edge detection module (▬ in Fig. 2 (b)) is proved in §3.2. Finally, in §3.3, we present more implementation details.

3.1. Pyramid Attention Module

For each saliency network layer, a pyramid attention module is first incorporated to generate a more discriminative feature representation. In contrast to previous saliency models that treat all positions of saliency features equally, our model focuses on the features in important regions and considers multi-scale information. This is achieved using a stacked attention architecture: multiple attention layers built upon multi-scale features are stacked to form a unified pyramid attention model.

More technically, let $\mathbf{X}$ denote a 3D feature tensor from a convolution layer of a saliency network (▬ in Fig. 2 (c)). This typically consists of C channels of width M and height M: $\mathbf{X} \in \mathbb{R}^{M\times M\times C}$. Our goal is to learn a set of equally-spatial-sized attention masks that softly weight output saliency features $\mathbf{X}$ based on multi-scale information. Essentially, we obtain multi-scale features by gradually down-sampling $\mathbf{X}$ into multiple-resolutions $\{\mathbf{X}^n : \mathbf{X}^n \in \mathbb{R}^{\frac{M}{2^n}\times\frac{M}{2^n}\times C}, n=1,2,3,\ldots,N\}$ with N steps. For $\mathbf{X}^n$ within a certain scale n, we use a soft attention mechanism [48] that predicts an importance map $\mathbf{l} \in [0,1]^{\frac{M}{2^n}\times\frac{M}{2^n}}$. Specifically, a softmax operation is applied over $\frac{M}{2^n}\times\frac{M}{2^n}$ spatial locations. The location softmax can be thought of as the probability with which our model believes the corresponding region in the input feature is important. It is defined as:

$$\mathbf{l}_i^n = p(L=i|\mathbf{X}^n) = \frac{\exp(\mathbf{W}_i^n\mathbf{X}_i^n)}{\sum_{j=1}^{\frac{M}{2^n}\times\frac{M}{2^n}}\exp(\mathbf{W}_j^n\mathbf{X}_i^n)}, \tag{1}$$

where $i \in 1,\ldots,\frac{M}{2^n}\times\frac{M}{2^n}$, $\mathbf{W}_i^n$ are the weights of the hidden layer that maps to the i-th element of the location softmax, L is a random variable which can take 1-of-$\frac{M}{2^n}\times\frac{M}{2^n}$ values. $\mathbf{l}$ is the attention map, where $\sum_i^{\frac{M}{2^n}\times\frac{M}{2^n}}\mathbf{l}_i = 1$. Through the operations above, our model learns a normalized importance weight (attention map) for each region at a certain scale (▭ in Fig. 2 (c)). This is essential for saliency representation since salient areas should have higher weights.

Once the attention probabilities $\{\mathbf{l}^n\}_{n=1}^N$ over all

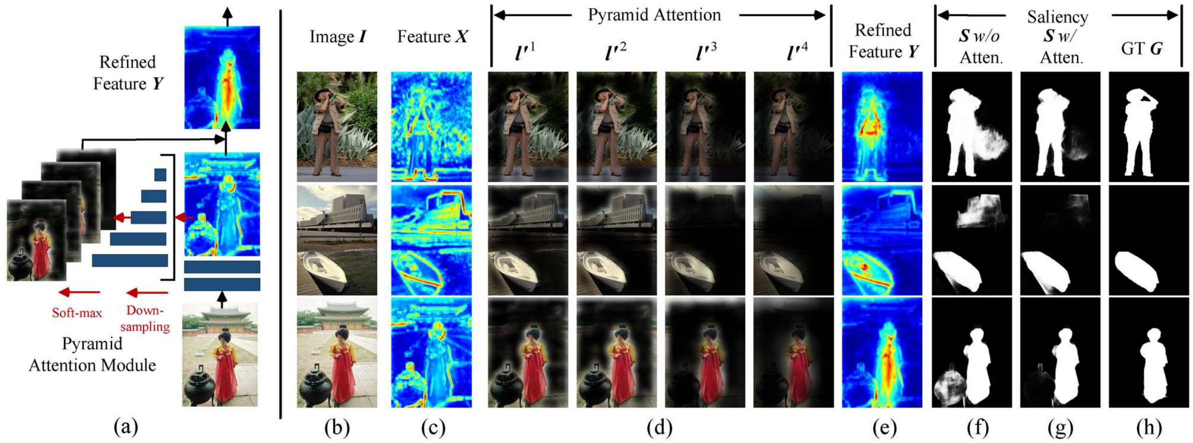


Figure 3: **Illustration of our pyramid attention module.** (a) Shows the work-flow of our attention module. (d) Gives the attention hierarchy that captures multi-scale information and emphasizes important regions. Comparing the features in (c) and (e), we find that the background responses have been successfully suppressed by the attention module. (f) and (g) show the results before/after applying attention. It can be observed that the PAGE-Net generates more accurate results through the attention module. See §3.1 for more details.

$\{\mathbf{X}^n\}_{n=1}^N$ are obtained, upsampling operations are adopted to resize them to their original resolutions: $\{\mathbf{l'}^n \in [0,1]^{M\times M}\}_{n=1}^N$. Fig. 3 offers a more detailed illustration of our attention module. Clearly, these attention maps (Fig. 3 (d)) correspond to different resolutions and can reveal important regions. More importantly, the pyramid attention module is equipped with stacked pooling operations, dramatically improving the receptive field of the corresponding feature extraction layer.

After calculating these importance probabilities, the original feature representation $\mathbf{X}$ is improved by accounting for the expectation of the feature slices in different regions:

$$\mathbf{Y}_j = \frac{1}{N}\sum\nolimits_{n=1}^{N} \mathbf{l'}^n_j \mathbf{X}_j, \qquad j \in 1, \ldots, M\times M, \tag{2}$$

where $\mathbf{Y}$ is the updated feature and $\mathbf{Y}_j$ is the j-th slice of the feature cube. Here, the model computes the expected value of the inputs by taking the expectation over the image features in different regions. Our attention module not only serves to enhance saliency representations in a focused location, but also accounts for multi-scale information. As discussed in [33], the features refined by the attention map usually have a large number of values close to zero. Thus, a stack of many refined features makes back-propagation difficult. To solve this, we apply identity mapping [12] in Eq. 2:

$$\mathbf{Y}_j = \frac{1}{N}\sum\nolimits_{n=1}^{N} (1+\mathbf{l'}^n_j)\mathbf{X}_j, \quad j \in 1, \ldots, M\times M. \tag{3}$$

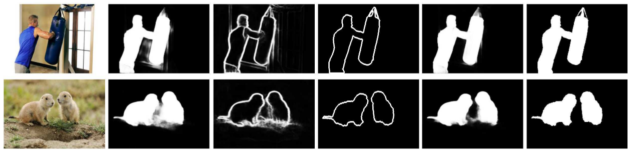
Even with a very small attention ($\mathbf{l'}_j \approx 0$), information from the original feature $\mathbf{X}$ will still be preserved by residual connection. As demonstrated in Fig. 3 (c) and (e), the attention module is able to enhance the feature map for more effective saliency representation. Such pyramid attention architecture provides a feasible method of assigning a global view of each corresponding convolution layer (with a significantly enlarged receptive field; see Fig. 2 (d)). A more detailed architecture of the attention module is presented in §3.3.

Discussion. Features from different positions do not contribute equally to saliency computation. Hence, we introduce the attention mechanism to focus on those positions most essential to the nature of salient objects. With our design, the attention module can quickly collect multi-scale information by iteratively downsampling the feature maps. Such a pyramid structure enables the receptive field of the feature layer to be easily and rapidly enlarged. Compared to previous attentive models, our pyramid attention is more favorable due to its effective use of multi-scale features and powerful representations with enlarged receptive fields, all of which are essential for pixel-wise saliency estimation.

3.2. Salient Edge Detector

With the refined saliency features $\mathbf{Y}$, a saliency map can be generated by directly feeding $\mathbf{Y}$ into a small stack of convolution layers with *sigmoid*, as done in previous methods. However, we observed that the detection cannot produce a clear boundary between the salient objects and the background (see Fig. 4 (b)). This is mainly due to the smoothness of the convolution kernel and the downsampling of the pooling layers. To deal with this, we design an extra salient edge detection module (see Fig. 2 (d)) to force the network to emphasize the saliency boundary alignment and learn to refine saliency maps with the use of salient edge information.

Let $\{(\mathbf{I}_k, \mathbf{G}_k, \mathbf{P}_k)\}_{k=1}^K$ denote the training data, where $\mathbf{I}_k$, $\mathbf{G}_k$, and $\mathbf{P}_k$ are the color image, the corresponding ground truth saliency map and the salient object boundary map, respectively. Notice that the edge map P_k (Fig. 4 (d))



(a) Image *I* (b) Saliency *S w/o* Edge Module $\mathcal{F}$ (c) Detected Salient Edge *E* (d) GT Salient Edge *P* (e) Saliency *S w/* Edge Module $\mathcal{F}$ (f) GT Saliency *G*

Figure 4: **Illustration of salient edge detection module of PAGE-Net.** The detected salient object edges in (c) offer important information on the location of salient objects. With this salient edge information, PAGE-Net is able to generate more accurate and better boundary-adherent results (e), compared with (b). See § 3.2 for more details.

can easily be obtained from the ground truth saliency map $\mathbf{G}_k$ (Fig. 4 (f)). We first build a salient edge detection module $\mathcal{F}(\mathbf{Y}_{I_k})$ (🟨 in Fig. 2 and Fig. 4 (c)), which can generate an estimated salient edge map (▭ in Fig. 2) for an input image $\mathbf{I}_k$. Here $\mathcal{F}$ denotes the salient edge detection module consisting of a stack of convolution layers and $\mathbf{Y}_{I_k}$ corresponds to the enhanced feature of $\mathbf{I}_k$. $\mathcal{F}$ can be learned by minimizing the following L2 norm loss function:

$$\begin{aligned}&\frac{1}{K}\sum\nolimits_{k=1}^{K}\mathcal{L}^{Edg}(\mathbf{P}_k,\mathcal{F}(\mathbf{Y}_{I_k})),\\ \mathcal{L}^{Edg}(\mathbf{P}_k,\mathcal{F}(\mathbf{Y}_{I_k}))&=||\mathbf{P}_k-\mathcal{F}(\mathbf{Y}_{I_k})||_2^2.\end{aligned}\tag{4}$$

A saliency readout network $\mathcal{R}(\mathbf{Y}_{I_k},\mathcal{F}(\mathbf{Y}_{I_k}))$ is then built to generate the saliency estimate (▭ in Fig. 2) by accounting for both saliency features $\mathbf{Y}_{I_k}$ and salient edge information $\mathcal{F}(\mathbf{Y}_{I_k})$. Thus the whole module can be learned by minimizing the following combination loss:

$$\frac{1}{K}\sum\nolimits_{k=1}^{K}\Big(\mathcal{L}^{Sal}(\mathbf{G}_k,\mathcal{R}(\mathbf{Y}_{I_k},\mathcal{F}(\mathbf{Y}_{I_k})))+\mathcal{L}^{Edg}(\mathbf{P}_k,\mathcal{F}(\mathbf{Y}_{I_k}))\Big),\tag{5}$$

where the saliency loss $\mathcal{L}^{Sal}$ is a weighted cross-entropy loss that accounts for data imbalance between salient and non-salient pixels:

$$\begin{aligned}\mathcal{L}^{Sal}\big(\mathbf{G},\mathcal{R}(\mathbf{Y}_I,\mathcal{F}(\mathbf{Y}_I))\big)=-\sum\nolimits_{i}&\beta(1-G_i)log(1-S_i)\\ &+(1-\beta)G_i log(S_i),\end{aligned}\tag{6}$$

where $i\in\Omega_I$, and Ω_I is the lattice domain of image $\mathbf{I}$. $\mathbf{S}$ indicates the saliency estimate for $\mathcal{R}$ and $S_i\in\mathbf{S}$. β refers to the ratio of salient pixels in the ground truth $\mathbf{G}$. With the loss function in Eq. 5 and the salient edge detection module $\mathcal{F}$, the readout network $\mathcal{R}$ learns to optimize the salient object estimates by leveraging explicit edge information.

Due to the hierarchical nature of the neural network, we introduce dense connection [16] to our model to make use of the information from different layers and increase representational ability. The saliency feature $\mathbf{Y}^{\ell}$ in the ℓ-th layer is enhanced by considering all multi-layer saliency estimates $\{\mathbf{S}^{\ell-1},\ldots,\mathbf{S}^1\}$, as well as edge information $\{\mathbf{E}^{\ell-1},\ldots,\mathbf{E}^1\}$ from all preceding $\ell-1$ layers:

$$\mathbf{Y}^{\ell}\leftarrow[\mathbf{Y}^{\ell},\mathcal{H}^{\ell}(\mathbf{E}^{\ell-1},\ldots,\mathbf{E}^1,\mathbf{S}^{\ell-1},\ldots,\mathbf{S}^1)],\tag{7}$$

where $\mathcal{H}$ indicates a small network that upsamples and concatenates the additional inputs from all preceding layers. Detailed architectures of $\mathcal{F},\mathcal{R},\mathcal{H}$ can be found in § 3.3.

Discussion. To preserve more boundary information, we add a salient edge detection module $\mathcal{F}$ that specifically focuses on segmenting salient object boundaries under the supervision of the ground truth edge map $\mathbf{P}$. Notice that $\mathcal{F}$ is general enough to incorporate other edge-aware filters like [6]. A readout network $\mathcal{R}$ for detecting salient objects is then learned using both the saliency feature $\mathbf{Y}$ and explicit salient edge information from $\mathcal{F}$. Dense connection is further introduced to draw representational power by reusing information from other layers.

3.3. Detailed Network Architecture

Backbone Network. The backbone network is built from the VGG-16 [31] model, which is well known for its elegance and simplicity and is widely used in saliency models. The first five convolutional blocks of VGG-16 are adopted. As shown in Fig. 5, we omit the last pooling layer (*pool5*) to preserve more spatial information.

Pyramid Attention Module. Let $\{\mathbf{X}^5,\mathbf{X}^4,\mathbf{X}^3,\mathbf{X}^2,\mathbf{X}^1\}$ denote the features from the last convolution layers of five conv blocks: *conv1-2*, *conv2-2*, *conv3-3*, *conv4-3*, and *conv5-3*. For each $\mathbf{X}^{\ell}$, we first downsample $\mathbf{X}^{\ell}$ into multiple scales. For scale n, the attention module is defined over three consecutive operations: BN→Conv($1\times1,1$)→ReLU, where the smallest attention map is set to 14×14. Upsampling operation is applied to resize the attention maps $\{\mathbf{l}^n\}_n$ over all scales to their original size. Then we obtain an enhanced saliency representation $\mathbf{Y}^{\ell}$ through Eq. 3.

Edge Detection Module. The edge detection module $\mathcal{F}$ is defined as: BN→Conv($3\times3,64$)→ReLU→Conv($1\times1,1$)→*sigmoid*. The saliency readout function $\mathcal{R}$ is built as: BN→Conv($3\times3,128$)→ReLU→BN→Conv($3\times3,64$)→ReLU→Conv($1\times1,1$)→*sigmoid*. For ℓ-th layer, a set of upsampling operations ($\mathcal{H}^{\ell}$) is adopted in order to enlarge all salient object estimations and salient edge information from all preceding layers with current feature resolutions. We then update the saliency representation $\mathbf{Y}^{\ell}$ through Eq. 7. Next, the edge detection module $\mathcal{F}$ and saliency readout function $\mathcal{R}$ are adopted to generate the corresponding saliency map $\mathbf{S}^{\ell}$.

Take *conv3-3* layer as an example. Given an input image $\mathbf{I}\in\mathbb{R}^{224\times224\times3}$, the saliency maps $\mathbf{S}^2,\mathbf{S}^1$ and edge maps $\mathbf{E}^2,\mathbf{E}^1$ from *conv4-3* and *conv5-3* layers are first upsampled into the current spatial resolution 56×56. Then are then fed into $\mathcal{H}^3$ and feature $\mathbf{Y}^3$ is updated accordingly. After applying the edge detection module $\mathcal{F}^3$ and saliency readout function $\mathcal{R}^3$, we obtain a saliency map $\mathbf{S}^3\in[0,1]^{56\times56}$. In this way, we get five saliency maps $\{\mathbf{S}^5,\mathbf{S}^4,\mathbf{S}^3,\mathbf{S}^2,\mathbf{S}^1\}$ from *conv1-2*, *conv2-2*, *conv3-3*, *conv4-3*, and *conv5-3*, respectively, where $\mathbf{S}^5\in[0,1]^{224\times224}$ is the final, most accu-

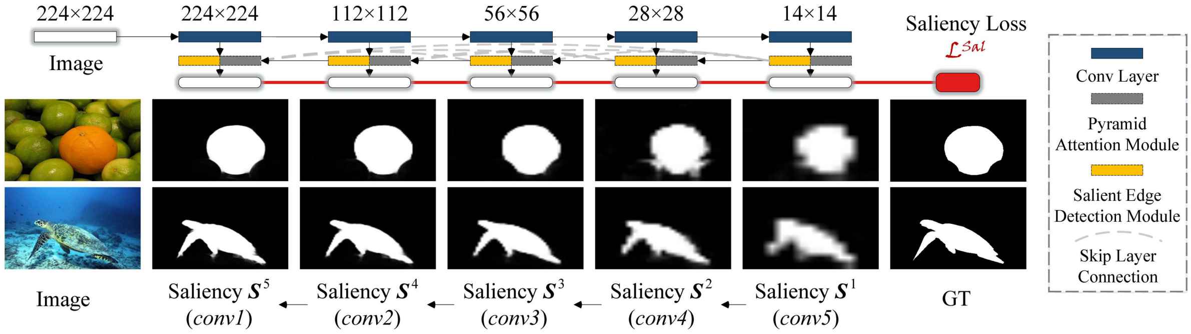
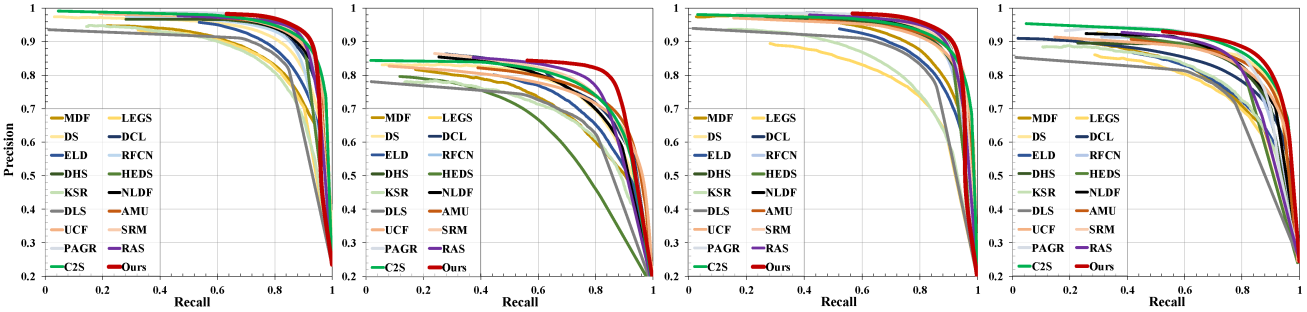


Figure 5: **Illustration of side outputs of PAGE-Net.** For better visualization, we omit the salient edge results. It can be observed that the saliency from different convolution blocks of VGG-16 can be gradually optimized in a top-down manner. See § 3.3 for details.



(a) ECCSD (b) DUT-OMRON (c) HKU-IS (d) PASCAL-S

Figure 6: **Quantitative results with PR-curve on four widely used benchmarks**: ECCSD [49], DUT-OMRON [50], HKU-IS [21] and PASCAL-S [25]. PAGE-Net gains promising performance. Best viewed in color. See § 4.1 for details.

rate saliency estimate.

Overall Loss. All the training images $\{\mathbf{I}_k\}_{k=1}^{K}$ are resized to fixed dimensions of $224\times224\times3$. The salient boundary maps $\mathbf{P}_k\in\{0,1\}^{224\times224}$ are generated from the corresponding ground truth salient object map $\mathbf{G}_k\in\{0,1\}^{224\times224}$ and dilated to a three-pixel radius. Considering all five-side outputs, the overall training loss for a training image $\mathbf{I}_k$ is:

$$\sum_{\ell=1}^{5}\Big(\mathcal{L}^{Sal}(\mathbf{G}_k^{\ell},\mathcal{R}^{\ell}(\mathbf{Y}_{I_k}^{\ell},\mathcal{F}^{\ell}(\mathbf{Y}_{I_k}^{\ell}))) + \mathcal{L}^{Edg}(\mathbf{P}_k^{\ell},\mathcal{F}^{\ell}(\mathbf{Y}_{I_k}^{\ell}))\Big). \tag{8}$$

With the hierarchical loss functions, five intermediate layers in PAGE-Net have direct access to the gradients from the loss function, leading to implicit deep supervision [19].

Implementation Details. PAGE-Net is implemented in Keras. Following the training protocol in [54, 20, 36], we use THUS10K [9], containing 10,000 images with pixel-wise annotations, for training. During the training phase, the learning rate is set to 0.0001 and is decreased by a factor of 10 every two epochs. In each training iteration, we use a mini-batch of 10 images. The entire training procedure takes about 7 hours using an Nvidia TITAN X GPU. Since our model does not need any pre- or post-processing, the inference only takes **0.04s** to process an image of size 224×224. This makes it faster than most deep learning based competitors (see § 4.1 for a detailed comparison).

4. Experiments

We conduct extensive experiments on six popular benchmarks: ECCSD [49], DUT-OMRON [50], HKU-IS [21], PASCAL-S [25], SOD [30], and DUTS-TE [35], which are all publicly available and are human-labeled with pixel-wise ground truth for quantitative evaluations. For evaluation, we adopt three widely used metrics [11], *i.e.*, precision-recall (PR) curves, F-measure and mean absolute error (MAE).

4.1. Performance Comparison

We compare the proposed PAGE-Net against 19 recent deep learning based alternatives: MDF [21], LEGS [34], DS [24], DCL [22], ELD [20], MC [57], RFCN [36], DHS [26], HEDS [14], KSR [38], NLDF [29], DLS [15], AMU [54], UCF [55], SRM [37], FSN [8], PAGR [56], RAS [7] and C2S [23]. we use either the implementations with the recommended parameter settings or the saliency maps shared by the authors. For a fair comparison, we exclude other ResNet-based models such as [39], or the ones using more training data [40]. Since fully connected conditional random field (CRF) has been used in [22, 14] as post-processing, we further offer a baseline PAGE-Net+CRF that uses CRF.

Methods	ECCSD [49]		DUT-OMRON [50]		HKU-IS [21]		PASCAL-S [25]		SOD [30]		DUTS-TE [35]	
	F-score ↑	MAE ↓	F-score ↑	MAE ↓	F-score ↑	MAE ↓	F-score ↑	MAE ↓	F-score ↑	MAE ↓	F-score ↑	MAE ↓
MDF* [21]	0.831	0.108	0.694	0.092	0.860	0.129	0.764	0.145	0.785	0.155	0.657	0.114
LEGS [34]	0.831	0.119	0.723	0.133	0.812	0.101	0.749	0.155	0.691	0.197	0.611	0.137
DS [24]	0.810	0.160	0.603	0.120	0.848	0.078	0.818	0.170	0.781	0.150	-	-
DCL [22]	0.898	0.071	0.732	0.087	0.907	0.048	0.822	0.108	0.784	0.126	0.742	0.150
ELD [20]	0.865	0.080	0.700	0.092	0.844	0.071	0.767	0.121	0.760	0.154	0.697	0.092
MC [57]	0.822	0.107	0.702	0.088	0.781	0.098	0.721	0.147	-	-	-	-
RFCN [36]	0.898	0.109	0.701	0.111	0.895	0.089	0.827	0.118	0.805	0.161	0.752	0.090
DHS* [26]	0.905	0.061	-	-	0.892	0.052	0.820	0.091	0.793	0.127	0.799	0.065
HEDS [14]	0.915	0.053	0.714	0.093	0.913	0.040	0.830	0.112	0.802	0.126	0.796	0.057
KSR [38]	0.801	0.133	0.742	0.157	0.759	0.120	0.649	0.137	0.698	0.199	0.660	0.123
NLDF [29]	0.905	0.063	0.753	0.080	0.902	0.048	0.831	0.112	**0.808**	0.130	0.777	0.066
DLS [15]	0.825	0.090	0.714	0.093	0.806	0.072	0.719	0.136	-	-	-	-
AMU [54]	0.889	0.059	0.733	0.097	**0.918**	0.052	0.834	0.103	0.773	0.145	0.750	0.085
UCF [55]	0.868	0.078	0.713	0.132	0.905	0.074	0.771	0.128	0.776	0.169	0.742	0.117
SRM [37]	0.910	0.056	0.707	0.069	0.892	0.046	0.783	0.127	0.792	0.132	0.798	0.059
FSN [8]	0.910	0.053	0.741	0.073	0.895	0.044	0.827	0.095	0.781	0.127	0.761	0.066
PAGR [56]	0.904	0.061	-	-	0.897	0.048	0.815	0.094	-	-	-	-
RAS* [7]	0.908	0.056	**0.758**	0.068	0.900	0.045	0.804	0.105	**0.809**	0.124	0.807	0.059
C2S [23]	0.902	0.054	0.731	0.080	0.887	0.046	0.834	0.082	0.786	0.124	0.783	0.062
PAGE-Net	**0.924**	**0.042**	**0.770**	**0.066**	**0.918**	**0.037**	**0.835**	**0.078**	0.796	**0.110**	**0.815**	**0.051**
PAGE-Net+CRF	**0.926**	**0.035**	**0.770**	**0.063**	**0.920**	**0.030**	**0.835**	**0.074**	0.796	**0.108**	**0.817**	**0.047**

*DHS [26] uses THUS10K and DUT-OMRON for training. MDF [21] and RAS [7] are trained on a subset of HKU-IS.

Table 1: **Quantitative results with F-measure (higher is better) and MAE (lower is better) on six well-known SOD benchmarks**: ECCSD [49], DUT-OMRON [50], HKU-IS [21], PASCAL-S [25], SOD [30] and DUTS-TE [35]. For each column, the top two best entries are highlighted in **red** and **blue**, respectively. See § 4.1 for details.

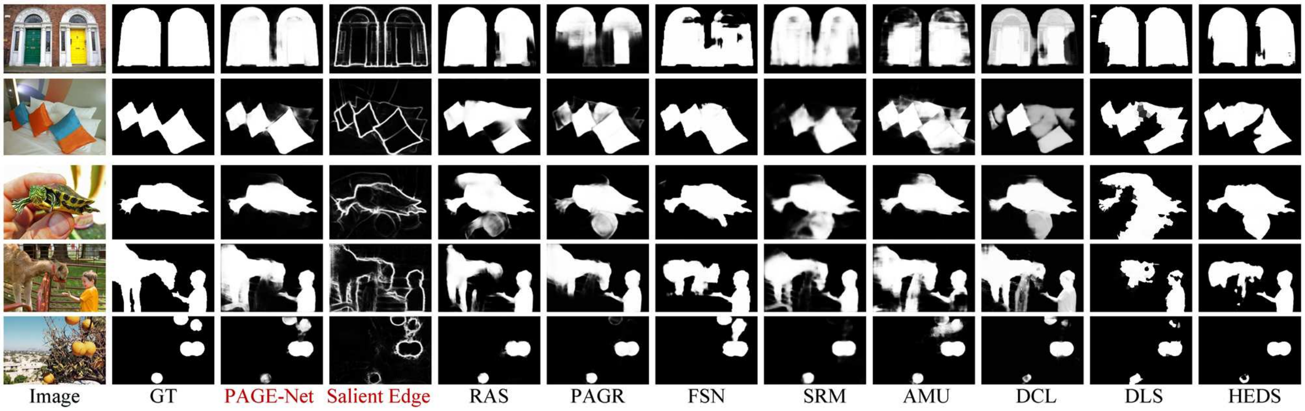


Figure 7: **Quantitative comparison of visual results on some representative challenging examples.** It can be observed that the proposed PAGE-Net is able to handle diverse challenging scenes. Best viewed in color. See § 4.1 for details.

Quantitative Evaluation. The precision-recall curves of all methods are given in Fig. 6. Due to limited space, we only show the results on four datasets. As seen, our PAGE-Net outperforms its counterparts across all datasets, convincingly demonstrating the effectiveness of the method. We also compare our method to current state-of-the-art models in terms of F-measure and MAE scores. It is evident from Table 1 that PAGE-Net achieves excellent results for all the datasets, across the metrics. In particular, PAGE-Net shows a significantly improved F-measure compared to the second best method, RAS, for the DUT-OMRON dataset (0.770 *vs* 0.758), which is one of the most challenging benchmarks. This clearly demonstrates the superior performance of PAGE-Net in complex scenes.

Qualitative Evaluation. Fig. 7 shows a visual comparison of the results of our method against those of five other top-

Method	LEGS [34]	MDF [21]	DS [24]	DCL [22]	ELD [20]
Time(s)	1.54	7.83	0.13	0.39	0.55
Method	RFCN [36]	DHS [26]	HEDS [14]	KSR [38]	NLDF [29]
Time(s)	4.65	0.04	0.57	49.64	0.09
Method	DLS [15]	AMU [54]	UCF [55]	SRM [37]	**PAGE-Net**
Time(s)	0.08	0.07	**0.04**	0.07	**0.04**

Table 2: **Runtime comparison** (GPU time) with previous deep learning based saliency models. See § 4.1 for details.

performing competitors. For better visualization, we highlight the main difficulties of each image group. We find that PAGE-Net performs well in a variety of challenging scenarios, *e.g.*, for large salient objects (first row), low contrast between objects and backgrounds (second row), cluttered backgrounds (forth row), and multiple disconnected objects (last row). Additionally, we observe that our method captures salient boundaries quite well due to its use of salient edge detection modules.

Runtime Comparison. We also report the runtime of several deep saliency methods in Table 2. These evaluations were conducted on a machine with an i7 CPU and a Titan-X GPU. PAGE-Net is faster than most of the others methods, achieving a real-time speed of 25 FPS.

4.2. Ablation Studies

In this section, we analyze the contribution of each component to the model's overall performance. We conduct experiments using the ECCSD [49] and DUT-OMRON [50] datasets. The results are summarized in Table 3.

Multi-Scale Attention. To validate the effectiveness of our multi-scale attention structure (§ 3.1), we compare three variants: *w/o attention*, *w/ single scale* and *w/o identity mapping*. Baseline *w/o attention* refers to the results obtained by retraining PAGE-Net without any attention module. The baseline *w/ single scale* corresponds to the results obtained with a single-scale attention module ($N = 1$ in Eq. 3). For *w/o identity mapping*, we retrain our attention module without identity mapping (Eq. 2). As shown in Table 3, the network with multi-scale attention achieves better performance, compared to those without an attention module or using single-scale attention. This confirms that the attention module benefits from multi-scale information. These results additionally demonstrate that identify mapping also boosts performance. The visual comparison between the results of PAGE-Net *w/* and *w/o* an attention module can be found in Fig. 3 (f) and (g).

Salient Edge Information. Next, we study the effect of salient object edge information (§ 3.2). The baseline *w/o salient edge* is obtained by disabling our salient edge detection module. We observe a drop in performance (ECCSD: 0.042→0.054, DUT-OMRON: 0.066→0.074) when using MAE. This suggests that the salient edge information does indeed improve salient object segmentation. To provide deeper insight into the importance of salient edge information, we est the model again after replacing the salient edge detection module with two different edge detectors: HED [47] and the canny filter. We also observe a minor decrease in performance in both cases. This indicates that the use of salient edge information is crucial for obtaining better performance. This is because salient edges offer an informative cue for detecting and segmenting salient objects, rather than simply determining color or intensity changes.

Aspects	Methods	ECCSD [49]		DUT-OMRON [50]	
		F-score ↑	MAE ↓	F-score ↑	MAE ↓
Full Model	PAGE-Net *conv 1*-output	**0.924**	**0.042**	**0.770**	**0.066**
Side Outputs	*conv 2*-output	0.914	0.051	0.764	0.070
	conv 3-output	0.906	0.056	0.761	0.072
	conv 4-output	0.887	0.068	0.740	0.083
	conv 5-output	0.854	0.090	0.706	0.099
Pyramid Attention Module	*w/o* attention	0.897	0.059	0.706	0.080
	w/ single scale	0.901	0.057	0.720	0.078
	w/o identity mapping (Eq. 2)	0.916	0.051	0.755	0.071
Salient-Edge Detection Module	*w/o* salient edge	0.910	0.054	0.746	0.074
	w/ HED [47]	0.911	0.052	0.751	0.073
	w/ canny detector	0.907	0.053	0.748	0.073

Table 3: **Ablation study of PAGE-Net** on ECCSD [49] and DUT-OMRON [50]. We change one component at a time, to assess individual contributions. See § 4.2 for details.

Side Outputs. Finally, we study the effect of our hierarchical architecture on inferring saliency in a top-down manner (Fig. 2 (b) and § 3.3). We introduced four additional baselines corresponding to the outputs from the intermediate layers of PAGE-Net: *conv2-output*, *conv3-output*, *conv4-output*, and *conv5-output*. Note that the final prediction of PAGE-Net can be viewed as the output from the $conv1$ layer. We find that the saliency results are gradually optimized by adding more details from the lower layers.

5. Conclusion

In this paper, we presented a novel deep saliency model, PAGE-Net, for salient object detection. PAGE-Net is equipped with two essential components: a pyramid attention module and a salient edge detection module. The former extends the regular attention mechanisms with multi-scale information to improve saliency representation, enabling more efficient training and better performance. The latter emphasizes on the detection of salient edge information, which can be leveraged for sharpening salient object segments. Extensive experimental evaluations over six well-known benchmark datasets verify that the aforementioned contributions significantly improve the saliency detection performance. Finally, the proposed model enjoys efficient inference speed and runs fast on GPU in real-time.

References

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

[2] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. Measuring the objectness of image windows. *IEEE TPAMI*, 34(11):2189–2202, 2012. 1

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015. 3

[4] Gedas Bertasius, Jianbo Shi, and Lorenzo Torresani. Semantic segmentation with boundary neural fields. In *CVPR*, 2016. 2

[5] Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang, Chang Huang, Wei Xu, et al. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *ICCV*, 2015. 3

[6] Liang-Chieh Chen, Jonathan T Barron, George Papandreou, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform. In *CVPR*, 2016. 2, 5

[7] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. Reverse attention for salient object detection. In *ECCV*, 2018. 3, 6, 7

[8] 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. 6, 7

[9] 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, 6

[10] Runmin Cong, Jianjun Lei, Huazhu Fu, Qingming Huang, Xiaochun Cao, and Chunping Hou. Co-saliency detection for rgbd images based on multi-constraint feature matching and cross label propagation. *IEEE TIP*, 27(2):568–579, 2018. 2

[11] 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. 6

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[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 Torr. Deeply supervised salient object detection with short connections. In *CVPR*, 2017. 1, 2, 6, 7, 8

[15] Ping Hu, Bing Shuai, Jun Liu, and Gang Wang. Deep level sets for salient object detection. In *CVPR*, 2017. 2, 6, 7, 8

[16] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. In *CVPR*, 2017. 5

[17] 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, 2

[18] Jason Kuen, Zhenhua Wang, and Gang Wang. Recurrent attentional networks for saliency detection. In *CVPR*, 2016. 2

[19] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *AISTATS*, 2015. 6

[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, 7, 8

[21] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *CVPR*, 2015. 2, 6, 7, 8

[22] Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In *CVPR*, 2016. 2, 6, 7, 8

[23] Xin Li, Fan Yang, Hong Cheng, Wei Liu, and Dinggang Shen. Contour knowledge transfer for salient object detection. In *ECCV*, 2018. 2, 6, 7

[24] Xi Li, Liming Zhao, Lina Wei, Ming-Hsuan Yang, Fei Wu, Yueting Zhuang, Haibin Ling, and Jingdong Wang. Deepsaliency: Multi-task deep neural network model for salient object detection. *IEEE TIP*, 25(8):3919 – 3930, 2016. 6, 7, 8

[25] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *CVPR*, 2014. 2, 6, 7

[26] Nian Liu and Junwei Han. DHSNet: Deep hierarchical saliency network for salient object detection. In *CVPR*, 2016. 2, 6, 7, 8

[27] Nian Liu, Junwei Han, and Ming-Hsuan Yang. PiCANet: Learning pixel-wise contextual attention for saliency detection. In *CVPR*, 2018. 3

[28] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. In *CVPR*, 2007. 2

[29] Zhiming Luo, Akshaya Mishra, Andrew Achkar, Justin Eichel, Shaozi Li, and Pierre-Marc Jodoin. Non-local deep features for salient object detection. In *CVPR*, 2017. 2, 6, 7, 8

[30] Vida Movahedi and James H Elder. Design and perceptual validation of performance measures for salient object segmentation. In *CVPR - Workshops*, 2010. 2, 6, 7

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

[32] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper convlstm for video salient object detection. In *ECCV*, 2018. 2

[33] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *CVPR*, 2017. 1, 3, 4

[34] Lijun Wang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Deep networks for saliency detection via local estimation and global search. In *CVPR*, 2015. 1, 2, 6, 7, 8

[35] 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. 2, 6, 7

[36] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Xiang Ruan. Saliency detection with recurrent fully convolutional networks. In *ECCV*, 2016. 2, 6, 7, 8

[37] Tiantian Wang, Ali Borji, Lihe Zhang, Pingping Zhang, and Huchuan Lu. A stagewise refinement model for detecting salient objects in images. In *ICCV*, 2017. 2, 6, 7, 8

[38] Tiantian Wang, Lihe Zhang, Huchuan Lu, Chong Sun, and Jinqing Qi. Kernelized subspace ranking for saliency detection. In *ECCV*, 2016. 6, 7, 8

[39] 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. 6

[40] Wenguan Wang, Jianbing Shen, Xingping Dong, Ali Borji, and Ruigang Yang. Inferring salient objects from human fixations. *IEEE PAMI*, 2019. 2, 6

[41] Wenguan Wang, Jianbing Shen, and Haibin Ling. A deep network solution for attention and aesthetics aware photo cropping. *IEEE TPAMI*, 2018. 1

[42] Wenguan Wang, Jianbing Shen, and Fatih Porikli. Saliency-aware geodesic video object segmentation. In *CVPR*, 2015. 1

[43] Wenguan Wang, Jianbing Shen, Ling Shao, and Fatih Porikli. Correspondence driven saliency transfer. *IEEE TIP*, 25(11):5025–5034, 2016. 2

[44] Wenguan Wang, Jianbing Shen, Hanqiu Sun, and Ling Shao. Video co-saliency guided co-segmentation. *IEEE TCSVT*, 28(8):1727–1736, 2018. 1

[45] Wenguan Wang, Jianbing Shen, Yizhou Yu, and Kwan-Liu Ma. Stereoscopic thumbnail creation via efficient stereo saliency detection. *IEEE TVCG*, 23(8):2014–2027, 2017. 1

[46] Yichen Wei, Fang Wen, Wangjiang Zhu, and Jian Sun. Geodesic saliency using background priors. In *ECCV*, 2012. 2

[47] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, 2015. 8

[48] 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*, 2015. 3

[49] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *CVPR*, 2013. 1, 2, 6, 7, 8

[50] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013. 2, 6, 7, 8

[51] Jimei Yang, Brian Price, Scott Cohen, Honglak Lee, and Ming-Hsuan Yang. Object contour detection with a fully convolutional encoder-decoder network. In *CVPR*, 2016. 2

[52] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, 2016. 3

[53] Jing Zhang, Yuchao Dai, Fatih Porikli, and Mingyi He. Deep edge-aware saliency detection. *arXiv preprint arXiv:1708.04366*, 2017. 2

[54] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *ICCV*, 2017. 2, 6, 7, 8

[55] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Baocai Yin. Learning uncertain convolutional features for accurate saliency detection. In *ICCV*, 2017. 2, 6, 7, 8

[56] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. Progressive attention guided recurrent network for salient object detection. In *CVPR*, 2018. 3, 6, 7

[57] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Saliency detection by multi-context deep learning. In *CVPR*, 2015. 1, 2, 6, 7

[58] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. Saliency optimization from robust background detection. In *CVPR*, 2014. 2