Saliency detection or foreground segmentation is a fundamental and important task in computer vision, which can be treated as a pixel-wise classification problem. Recently, although fully convolutional network (FCN) based approaches have made remarkable progress in this task, segmenting salient objects in complex image scenes is still a challenging problem. In this paper, we argue that, when predicting the saliency of a given pixel, human-like attention mechanisms play an important role in structural saliency inference. Therefore, we propose a simple yet surprisingly effective self-gated soft-attention mechanism for fast saliency detection. The soft-attention mechanism generates a gating signal that is end-to-end trainable, which allows deep networks to contextualize local information useful for saliency prediction. In addition, the proposed attention mechanism is channel-wise, generic and can be easily incorporated into any existing FCN architectures like Trojan Horse, while only requiring negligible parameters. Extensive experiments verify the superior effectiveness of the proposed method. More specifically, our method achieves a new state-of-the-art performance on seven public saliency benchmarks, and outperforms the very recent methods with a large margin.

1 Introduction

When human observers look at an image, effective attentional mechanisms attract their gazes on salient regions which have distinctive variations in visual stimuli. Emulating such ability has been studied for more than 90 years by neuroscientists [4] and more recently by computer vision researches. Traditionally, saliency prediction algorithms focus on identifying the eye fixation points that human observers would focus on at first glance. In recent years, due to large-scale scalable applications, more and more efforts of saliency research have been made to automatically identify and segment the most conspicuous objects or regions in an image. In general, saliency prediction is regarded as the first step to narrow down subsequent object-related vision tasks. For example, saliency detection can be used in pattern classification [44], image retrieval [38], semantic segmentation [1, 6], image manipulation [26], visual tracking [25, 46] and person re-identification [51], etc.

Recently, with the prevalence of deep architectures, remarkable progresses have been achieved in a wide range of computer vision tasks supported by large-scale training datasets, such as image classification [13, 32], object detection [9], and semantic segmentation [22]. Thus, many researchers start to shift their efforts to deep convolutional neural networks (CNNs) and train large-scale CNNs to automatically extract powerful feature representations for saliency detection. Although effective, most of state-of-the-art methods require pre-trained CNNs, which usually employ the strided convolution and pooling operations. These downsampling methods increase the receptive field of CNNs, helping to extract high-level semantic features, nevertheless they inevitably drop the location information and fine details of objects, leading to unclear boundary predictions. Furthermore, the lack of structural supervision also makes saliency prediction a challenging problem in complex image scenes.

*The corresponding author. This work was done during the first author’s visit at the University of Adelaide.
Figure 1: Image examples of diverse scenes. If we only see the parts of the images, salient objects could be mistaken. However, if we can have a glance at the whole image even shorter than one second, it is easy to tell where salient objects are coarsely located depending on their surroundings. The original images are captured from the Westworld Series (https://www.hbo.com/westworld/season-01).

As for our human-beings, visual attention mechanism is fairly important for us to fast locate objects in images [4]. When facing a complex visual scene as shown in Fig. 1, we can efficiently locate regions of interest and analyze the scene by selectively processing subsets of visual input. Even we check the entire image with a glance, we could recognize where salient objects are. Based on this fact, visual attention is a hot topic in computer vision, neuroscience and deep learning area. Many works have proven that attention modules can significantly boost the performance. However, as will be discussed in the following section, existing attention modules only use a small context of feature descriptors to generate local attention masks, which will lose some important contextual information.

To remedy the aforementioned problems, we propose a novel attention module which can be easily incorporated into any existing deep architectures and learned in the end-to-end manner. To be specific, we design a self-gated soft-attention which consists of context-aware unit and soft-attention generator, as illustrated in Fig. 2 (b). The context-aware unit takes extracted features as inputs and performs large-context pooling to aggregate the complementary visual features for generating attention. The soft-attention generator exports channel-wise masks to integrate the local features for feature selection. In addition, to effectively train our network, we propose a novel loss function which incorporates structural information to supervise the training process. Equipped with these new schemes, our proposed model effectively mimics the visual attention of human and captures the boundaries and spatial contexts of salient objects, hence significantly boosts the saliency performance.

In summary, our contributions are three folds:

- We propose a novel self-gated soft-attention mechanism in the context of pixel-wise image classification. We apply the proposed attention model to saliency detection and show its superior classification performance over the baseline approaches.
- We demonstrate that the proposed attention mechanism can be incorporated into any existing deep architectures and provide fine-scale attention maps with minimal computational overhead, which is a crucial step towards explainable deep learning.
- Extensive experiments on seven large-scale saliency benchmarks demonstrate that the proposed approach achieves superior performance and outperforms the very recent state-of-the-art methods with a large margin.

2 Related Work

Visual Saliency Detection. Saliency detection is important in understanding the content of images and finding target objects. Over the past two decades, lots of saliency detection methods have been developed. The majority of existing methods are based on hand-crafted features. Most of them have been well summarized in [4]. Nowadays, thanks to the powerful feature representations, deep learning based approaches have achieved impressive performance improvements. Most, if not all, of the state-of-the-art saliency models are based on the fully convolutional network (FCN) [22]. For example, Wang et al. [36] develop a recurrent version of FCN to incorporate saliency priors for saliency map inference. Liu et al. [21] propose a deep hierarchical network to learn a coarse global estimation and then refine the saliency map, progressively. Hou et al. [10] naively introduce dense short connections to the holistically-nested edge detection (HED) architecture [39] to enrich multi-scale features for saliency prediction. Zhang et al. [47] propose to aggregate multi-level convolutional features for saliency detection. And they also develop a novel dropout method to learn uncertain
convolutional features to improve the robustness and accuracy of saliency detection [48]. Wang et al. [37] provide a stage-wise refinement model to gradually get accurate saliency detection results. Inspired by the image intrinsic reflection, Zhang et al. [45] design a symmetrical FCN to learn lossless complementary features for accurate saliency detection. To accelerate the speed of saliency detection, Zhang et al. [49] utilize contextual pyramids to fast locate salient objects, resulting real-time detection. Despite aforementioned approaches achieve remarkable success, there still exist several obvious problems. For example, due to the usage of strided pooling and convolution layers, the size of the predicted map is usually much smaller than that of the original image, which makes it difficult to assign labels for every pixel in the input image. To solve this issue, essential upsampling methods are needed to rescale the predicted map to the size of the input image. However, they inevitably drop the location information and fine details of objects, leading to unclear boundary predictions. In addition, the receptive field of commonly used pre-trained models are not large enough.

**Visual Attention Mechanism.** Attention mechanisms were first popularized in the context of natural language processing (NLP), especially in machine translation [24]. Basically, attention mechanisms can be separated into two types: soft-attention and hard-attention. In soft-attention, continuous functions (e.g., soft-max, sigmoid) are used to assign the attention weight on the input, making it fully differentiable. In contrast, hard-attention models propose specific regions by sampling from the weights. As the sampling operation is not differentiable, hard-attention is trained using the gradient of the likelihood term generated by Monte-Carlo sampling. In computer vision, attention mechanisms are applied to a variety of problems, including image classification [11, 33], action recognition [20, 29], image captioning [23, 40], and visual question answering [42, 28]. However, for pixel-wise classification tasks, despite the importance of local information, only a handful of works use attention mechanisms [30, 49]. In these methods, either extra labels are available to guide the attention, or the local context is extracted by a hard-attention model. In our work, we aim to improve the attention modules with feature aggregation approaches. We incorporate large-context information into the mask generation, deriving a self-gated soft-attention approach that is end-to-end trainable.

### 3 Gated Attention Network Model

Fig. 2 illustrates the semantic overview of our Gated Attention Network (GANet) for saliency detection. GANet is constructed by stacking multiple Self-Gated Soft-Attention modules (Fig. 2(b)) with its FCN backbone. Each attention module is divided into two branches: feature extraction branch and attention generation branch. The feature extraction branch performs multi-level feature extraction and can be adapted to any state-of-the-art FCN architectures. The attention generation branch learns attention maps which focus computation on specific parts of the input features relevant to the salient objects. In the following subsections, we elaborate core components of the proposed model.

#### 3.1 Existing Spatial Attentions

Before we talk about the proposed method, we need to take a detour and briefly review the existing attention methods. Generally, existing attention structures consist of two modular branches: the feature extraction branch and attention generation branch. The attention generation branch is jointly trained with the feature extraction branch end-to-end. Formally, we denote the output features of some convolutional/pooling layers by \( X \in \mathbb{R}^{W \times H \times C} \). The attention generator takes feature map \( X \) as the input and outputs spatial-normalized attention mask \( A \in \mathbb{R}^{W \times H} \). A is applied on \( X \) to get attended feature \( \hat{X} \in \mathbb{R}^{W \times H \times C} \). For notation simplicity, we denote the 3-D feature output of FCNs with upper-case letters and feature at one spatial location with its corresponding lower-case letters. For example, \( x_{i,j} \in \mathbb{R}^{C} \) is the feature at the position \((i, j)\) of \( X \). Attention generator outputs attention mask \( A \) which acts as a spatial regularizer to enhance the relevant regions and suppress the non-relevant regions for feature \( X \). The details of the attention generation branch are as follows: an attention generator consists of a convolutional layer, a non-linear activation layer and a spatial-normalization as

\[
\alpha_{i,j} = \frac{z_{i,j}}{\sum_{i,j} z_{i,j}} = \sigma(w^T x_{i,j} + b),
\]

where \( \sigma \) is a non-linear activation function, such as softmax and sigmoid function in [30, 49]. \( w_{i,j} \in \mathbb{R}^{C} \) and \( b \in \mathbb{R} \) are the parameters of the attention generator model, which is a \( 1 \times 1 \) convolution layer. The attended feature \( \hat{x}_{i,j} \in \mathbb{R}^{C} \) is calculated as \( \hat{x}_{i,j} = x_{i,j} \alpha_{i,j} \). This kind of spatial
Our proposed GANet framework.

(a) Our proposed GANet framework.

(b) Self-gated soft-attention module

Figure 2: Semantic illustration of our proposed GANet framework. We adopt the modified VGG-16 model [47] as its backbone and each convolutional block can be expanded to a soft attention block. Note that we display the channel dimension of feature maps. More details can be found in the text.

attentions is a single mask mechanism to assign different weights to different feature spatial regions depending on their feature content. It automatically predicts the weighted heat map to enhance the relevant features and suppress the irrelevant features during the training process for specific tasks.

3.2 Large Grid Pooling based Context Aware Unit

In previous works [30, 49], the attention masks are generated directly by sampling the original features with involving simple convolutional layers (Section 3.1). This results in a lazy context-free soft attention mask generator, which can yield satisfactory performance only when we do not need to localize which image parts should be attended in a suitable context. However, if visual structures are very complex and salient objects can be localized only in a certain context, the attention mask generator should be equipped with a frontend unit to model visual contexts. Inspired by the dilation convolutional structure [43], we introduce a novel context-aware unit. It is constructed by stacking a large-context pooling and convolutional layers together. This structure can be viewed as aggregating the input locally by applying a non-linearity to the content of some patches. More specifically, for the large-context pooling, we use the dilation grid pooling, which can be defined as

\[ x(p, q) = \text{pooling}(x_{i,j}, x_{i+p,j+p}, ..., x_{i+pq,j+pq}). \]  

(2)

where \( p \in \mathbb{N}^1 \) is a dilation factor and \( q \in \mathbb{N}^1 \) is the cover size. The proposed dilation grid pooling is the “pooling with a dilated stride” in the grid structure. The large-context pooling operator can apply the same pooling at different ranges using different dilation factors. Our definition uses the grid selection of the input features, which involves large-region contextual information. The rationalities can be summarized as follow: 1) when aggregating local and contextual information, we should consider as many descriptors as possible, which helps us to get comprehensive context features for pixel-wise classification. 2) although more descriptors are required, we should give different attention
to them because descriptors near the target pixel usually contain more related semantic information, and we are supposed to build a fine representation for it. For those descriptors far from the given pixel, a coarse representation should be enough. On the other hand, in the convolutional layers, we simply employ a kernel filter of size $1 \times 1$ to ensure the equal-size feature maps. Eventually, the output layer of this “context-aware unit” has the same size as the input features, but captures a larger size of receptive field and deeper context-aware features. In this fashion, attention masks can be generated side-by-side by inserting such contextual layer (see Fig. 2 and section 3.3). Some soft attention models (e.g., [33, 49]) also consider context-aware information to guide attention selection but with more complicated structures. In addition, with high complexity, their designs are ineffective in model deployment and prone to being overfitting when only a small set of labeled data is available for model training like pixel-wise classification tasks.

3.3 Soft Attention Generator

Based on the pooled contextual features in section 3.2, we describe the gated soft attention generator. Formally, suppose $P(X)$ is the output feature map of the context-aware unit with the input feature $X$. The soft attention generator derives an attention mask $M(X)$ of the same size with $P(X)$ through a channel-wise sigmoid modification. Thus, the attention-aware feature map can be computed as

$$X_{i,j,c}^a = M_{i,j,c}(X) \ast P_{i,j,c}(X),$$

where $i \in \{1, ..., W\}, j \in \{1, ..., H\}, c \in \{1, ..., C\}$ index the width, the height and the channel of the feature map, respectively. In our soft attention generator, the mask $M_{i,j,c}(X)$ is computed by:

$$M_{i,j,c}(X) = \frac{1}{1 + exp(-W_{\text{att}}A_{i,j,c}(X))}, A_{i,j,c}(X) = \text{ReLU}(W_{\text{con}} \ast P_{i,j,c}(X) + b_{\text{con}}),$$

Here, $W_{\text{att}}$ is the attention parameter. $W_{\text{con}}$ and $b_{\text{con}}$ are the context aggregation parameters. All of them would be learnt during the training process. Such structure extend the conventional attention structure to multi-channel cases. Moreover, it generates attention maps of different contexts which identify object regions and predict the discriminative parts jointly. In addition, the proposed gated attention mechanism can be directly embedded into any state-of-the-art FCN frameworks.

3.4 Hierarchical Feature Interaction

Conceptually, the above attention-aware feature learning aims at depicting the most discriminative local regions of input features, while the original feature $X$ is dedicated to encoding the optimal global level features from the entire image. In this sense, the attention-aware features can be viewed as some kind of local features and are largely complementary with the original features in functionality. Intuitively, their combination can integrate both advantages (i.e., preserving global information and being more sensitive to particular local positions) and relieve the modeling burden from the same (particularly small) training data. Therefore, we further introduce a hierarchical feature interaction scheme for maximizing the complementary benefit and compatibility of both the global and local feature representations. To be specific, the original features are combined with the obtained attention-aware features by dense skip connections. Formally, the interaction is defined by

$$\varphi_l(X) = \begin{cases} \psi([\phi_l(X), \varphi_{l+1}(X), \phi_l(X^a)]), l < L \\ \psi([\phi_{L}(X), \phi_{L}(X^a)]), l = L \end{cases}$$

where $\psi$ denotes the interaction operator, which is a $1 \times 1$ convolutional layer followed by a deconvolutional layer to ensure the same resolution. $[,]$ is the concatenation operator in channel-wise. $\phi_l$ is the feedforward process to extract features of the $l$-th convolutional block. For the saliency map prediction, we add a convolutional layer with two filters after the integrated features.

3.5 GANet Model Training

In this paper, we define a new loss function that handles the class imbalance problems in saliency detection. More specifically, our model is trained by optimizing two loss functions: (i) weighted binary cross-entropy loss (pixel based) and (ii) Dice loss (shape based). The cross-entropy loss provides a probabilistic measure of similarity between the prediction and ground truth. The Dice loss is inspired by the Dice overlap ratio and yields a true positive count based estimate of similarity [27].
After passing input images through the defined GANet, each pixel \( x \) will have a score estimation \( y_m(x) \) to belong to the class \( m \). Pixel-wise softmax is used to transform \( y_m(x) \) into the probability \( p_m(x) \in [0,1] \) by \( p_m(x) = \frac{e^{y_m(x)}}{\sum_{m=0}^{M} e^{y_m(x)}} \). The ground truth probability is \( g_m(x) \). Thus, the loss function is defined by

\[
L = -\alpha \sum_x w(x) g_m(x) \log(p_m(x)) - (1 - \alpha) \frac{2 \sum_x p_m(x) g_m(x)}{\sum_x p_m^2(x) + \sum_x g_m^2(x)},
\]

where \( \alpha \) controls the amount of loss terms. We introduce weights \( w(x) \) to tailor the loss function to challenges that we have encountered in saliency detection: the class imbalance and the segmentation errors along object boundaries. Given the frequency \( f_m \) of class \( m \) in the training data, \( \text{i.e., the class probability, the indicator function } I \), the training segmentation \( S \), and the 2D gradient operator \( \nabla \), the weights are defined as

\[
w(x) = \sum_m I(S(x) == m) \frac{\text{median}(f)}{f_m} + w_0 I(|\nabla S(x)| > 0),
\]

with the vector of all frequencies \( f = [f_0, ..., f_M] \). The first term models median frequency balancing \( 2 \) and compensates for the class imbalance problem by highlighting classes with low probability. The second term puts higher weight on object boundary regions to emphasize on the correct segmentation of contours. \( w_0 \) balances the two terms.

4 Experiments

4.1 Experimental Settings

Evaluation Datasets. For the performance evaluation, we adopt seven public saliency datasets described as follows: DUT-OMRON \( 41 \) dataset has 5,168 high quality natural images. Each image in this dataset has one or more objects with relatively complex image background. DUTS-TE dataset is the test set of currently largest saliency detection benchmark (DUTS) \( 35 \). It contains 5,019 images with high quality pixel-wise annotations. ECSSD \( 31 \) dataset contains 1,000 natural images, in which many semantically meaningful and complex structures are included. HKU-IS-TE \( 16 \) dataset has 1,447 images with pixel-wise annotations. Images of this dataset are well chosen to include multiple disconnected objects or objects touching the image boundary. PASCAL-S \( 19 \) dataset is generated from the PASCAL VOC \( 7 \) dataset and contains 850 natural images with segmentation-based masks. SED \( 3 \) dataset has two non-overlapped subsets, \( \text{i.e., SED1 and SED2}. \) SED1 has 100 images each containing only one salient object, while SED2 has 100 images each containing two salient objects. SOD \( 12 \) dataset has 300 images, in which many images contain multiple objects either with low contrast or touching the image boundary.

Evaluation Criteria. To evaluate our method with other algorithms, we follow the works in \( 47 \), \( 48 \), \( 45 \) and use four main metrics as the standard evaluation score (with typical parameter settings), including the widely used precision-recall (PR) curves, F-measure \( (\eta = 0.3) \), mean absolute error (MAE) and S-measure \( (\lambda = 0.5) \). More details can be found in \( 4 \), \( 8 \).

Implementation Details. We implement our proposed model based on the modified Caffe toolbox \( 2 \) with the MATLAB 2016 platform. We train and test our method with an NVIDIA Titan 1070 GPU (with 8G memory). For training our model, we follow the works in \( 47 \), \( 48 \), \( 45 \), and adopt the MSRA10K \( 4 \) dataset, which has 10,000 training images with high quality pixel-wise saliency annotations. We also augment this dataset by random cropping and mirror reflection, producing 120,000 training images totally. We do not use validation set and train the model until its training loss converges. The input image is uniformly resized into \( 384 \times 384 \times 3 \) pixels and subtracted the ImageNet mean \( 5 \). The weights of feature extraction parts are initialized from the VGG-16 model \( 32 \). For other layers, we initialize the weights by the “msra” method. To speed the training process, we add the batch normalization (BN) \( 2 \) before each ReLU layer. During the training, we use standard SGD method with a batch size 12, momentum 0.9 and weight decay 0.0005. We set the base learning rate to 1e-7 and decrease the learning rate by 10% when training loss reaches a flat. We found that the equal contribution \( (i.e., \alpha = 0.5, w_\eta = 1) \) of pixel based and shape based terms gives the best results. The training process took about 40 hours. When testing, our proposed model runs at about 12.2 fps. The source code will be made publicly available upon the acceptance of this work.
4.2 Comparison with State-of-the-Arts

We compare our method with other 12 state-of-the-art ones, including AMU [47], DCL [17], DHS [21], DS [18], DSS [10], ELD [15], LEGS [34], LFR [45], MCDL [50], MDF [16], RFCN [36], UCF [48]. For fair comparison, we use either the implementations with recommended parameter settings or the saliency maps provided by the authors.

Quantitative Evaluation. As illustrated in Tab. [1] Tab. [2] and Fig. [3], our method outperforms other competing ones across all datasets in terms of near all evaluation metrics. From these results, we have other notable observations: (1) our method consistently achieves higher S-measure than other methods, especially on complex structure datasets, e.g., the SED and SOD datasets. We attribute this result to our structural loss. (2) Without segmentation pre-training and conditional random field (CRF) post-processing, our method still achieves better results than the DCL, DS, DSS and RFCN, especially on the DUTS-TE, ECSSD and SED datasets, where our method achieves about 2% performance leap of F-measure and around 4% improvement of S-measure, as well as around 3% decrease in MAE compared with existing best methods. (3) compared to the DSS, our method is inferior on the DUT-OMRON datasets. However, our method ranks in the second place and is still very comparable.

Ablation Studies. To verify the effectiveness of each component in our model, we design several ablation studies with different settings on the ECSSD datasets, including different number of self-attention gates, w/ and w/o using large-context structure, different pooling strategies, different attention methods and different FCN backbones. Note that all unrelated settings are the same as the

| Models          | FCN baseline | Recurrent Attention | Residual Attention | Single Attention | Our methods |
|-----------------|--------------|--------------------|-------------------|-----------------|-------------|
| MAE             | 0.808        | 0.808              | 0.805             | 0.804           | 0.812       |
| SOD             | 0.815        | 0.815              | 0.814             | 0.813           | 0.816       |

Table 3: Results with different model settings on the ECSSD dataset.
Figure 3: The PR curves of different state-of-the-art methods. Our method is labeled as GAN.

Figure 4: Comparison of saliency maps. (a) Inputs; (b) GT; (c) Ours; (d) AMU; (e) DCL; (f) DHS; (g) DSS; (h) LFR; (i) MCDL; (j) RFCN; (k) UCF. We don’t show the results of DS, ELD, LEGS and MDF. More examples can be found in supplemental materials.

implementation details in section 4.1. Due to the limitation of space, we only show the results of different attention methods in the main text and list most of the results in supplemental materials. Tab. 3 shows the performance of our GANet with different attention methods. It can be observed that most of attention methods improve the performance, this further prove that attentive features do help the pixel-wise labeling tasks. The channel-wise residual attention [14] performs better than single mask attention [53, 49] in the most cases. The reason is that channel-wise attention takes all locations of a particular parts into account and all the locations contribute to the final representation. Experimental results also demonstrate that our method achieves much better results compared with using either of them. More visual comparisons are present in supplemental materials.

5 Conclusion

In this work, we propose a novel self-gated attention approach to address the challenging saliency detection task. The proposed attention exploits multi-level information of deep networks, which successfully selects the most related regions and enhances the discriminative ability of features for saliency detection. In addition, we leverage large-context pooling to mine the discriminative information of each part in a global-local manner. We also introduce a new structural loss function to supervise the model training. All the components detailed in this work can be easily embedded into any other FCN framework to make a further performance improvement. Extensive ablation studies and comparisons well demonstrate the effectiveness of our approach. In the future, we are planning to exploit more challenging pixel-wise classification tasks, such as semantic segmentation.
References

[1] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk. Salient region detection and segmentation. Computer Vision Systems, pages 66–75, 2008.

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. TPAMI, 39(12):2481–2495, 2017.

[3] A. Borji. What is a salient object? a dataset and a baseline model for salient object detection. IEEE TIP, 24(2):742–756, 2015.

[4] A. Borji, M.-M. Cheng, H. Jiang, and J. Li. Salient object detection: A benchmark. IEEE TIP, 24(12):5706–5722, 2015.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In CVPR, pages 248–255, 2009.

[6] M. Donoser, M. Urschler, M. Hirzer, and H. Bischof. Saliency driven total variation segmentation. In ICCV, pages 817–824, 2009.

[7] M. Everingham, L. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. IJCV, 88:303–338, 2010.

[8] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji. Structure-measure: A new way to evaluate foreground maps. In ICCV, pages 4548–4557, 2017.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, pages 580–587, 2014.

[10] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr. Deeply supervised salient object detection with short connections. In CVPR, pages 5300–5309, 2017.

[11] S. Jetley, N. A. Lord, N. Lee, and P. H. Torr. Learn to pay attention. arXiv:1804.02391, 2018.

[12] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li. Salient object detection: A discriminative regional feature integration approach. In CVPR, pages 2083–2090, 2013.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, pages 1097–1105, 2012.

[14] J. Kuen, Z. Wang, and G. Wang. Recurrent attentional networks for saliency detection. arXiv:1604.03227, 2016.

[15] G. Lee, Y.-W. Tai, and J. Kim. Deep saliency with encoded low level distance map and high level features. In CVPR, pages 660–668, 2016.

[16] G. Li and Y. Yu. Visual saliency based on multiscale deep features. In CVPR, pages 5455–5463, 2015.

[17] G. Li and Y. Yu. Deep contrast learning for salient object detection. In CVPR, pages 478–487, 2016.

[18] X. Li, L. Zhao, L. Wei, M.-H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang. Deepadsaliency: Multi-task deep neural network model for salient object detection. IEEE TIP, 25(8):3919–3930, 2016.

[19] Y. Li, X. Hou, C. Koch, J. Rehg, and A. Yuille. The secrets of salient object segmentation. In CVPR, pages 280–287, 2014.

[20] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot. Global context-aware attention lstm networks for 3d action recognition. In CVPR, pages 1647–1656, 2017.

[21] N. Liu and J. Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In CVPR, pages 678–686, 2016.

[22] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR, pages 3431–3440, 2015.

[23] J. Lu, C. Xiong, D. Parikh, and R. Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In CVPR, pages 375–383, 2017.

[24] M.-T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. arXiv:1508.04025, 2015.
[25] V. Mahadevan and N. Vasconcelos. Saliency-based discriminant tracking. In CVPR, pages 1007–1013, 2009.

[26] L. Marchesotti, C. Cifarelli, and G. Csurka. A framework for visual saliency detection with applications to image thumbnailing. In ICCV, pages 2232–2239, 2009.

[27] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In IC 3D Vision (3DV), pages 565–571, 2016.

[28] H. Nam, J.-W. Ha, and J. Kim. Dual attention networks for multimodal reasoning and matching. arXiv:1611.00471, 2016.

[29] W. Pei, T. Baltrušaitis, D. M. Tax, and L.-P. Morency. Temporal attention-gated model for robust sequence classification. In CVPR, pages 820–829, 2017.

[30] M. Ren and R. S. Zemel. End-to-end instance segmentation with recurrent attention. arXiv:1605.09410, 2017.

[31] J. Shi, Q. Yan, L. Xu, and J. Jia. Hierarchical image saliency detection on extended cssd. IEEE TPAMI, 38(4):717–729, 2016.

[32] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014.

[33] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. arXiv:1704.06904, 2017.

[34] L. Wang, H. Lu, X. Ruan, and M.-H. Yang. Deep networks for saliency detection via local estimation and global search. In CVPR, pages 3183–3192, 2015.

[35] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan. Learning to detect salient objects with image-level supervision. In CVPR, pages 136–145, 2017.

[36] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan. Saliency detection with recurrent fully convolutional networks. In ECCV, pages 825–841, 2016.

[37] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu. A stagewise refinement model for detecting salient objects in images. In ICCV, pages 4019–4028, 2017.

[38] W. Wang, Y. Song, and A. Zhang. Semantics-based image retrieval by region saliency. Image and Video Retrieval, pages 245–267, 2002.

[39] S. Xie and Z. Tu. Holistically-nested edge detection. In ICCV, pages 1395–1403, 2015.

[40] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In ICML, pages 2048–2057, 2015.

[41] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In CVPR, pages 3166–3173, 2013.

[42] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In CVPR, pages 21–29, 2016.

[43] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. arXiv:1511.07122, 2015.

[44] F. Zhang, B. Du, and L. Zhang. Saliency-guided unsupervised feature learning for scene classification. IEEE TGRS, 53(4):2175–2184, 2015.

[45] P. Zhang, W. Liu, H. Lu, and C. Shen. Salient object detection by lossless feature reflection. In IJCAI, pages --, 2018.

[46] P. Zhang, D. Wang, H. Lu, and H. Wang. Non-rigid object tracking via deep multi-scale spatial-temporal discriminative saliency maps. arXiv:1802.07957, 2018.

[47] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In ICCV, pages 202–211, 2017.

[48] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin. Learning uncertain convolutional features for accurate saliency detection. In ICCV, pages 212–221, 2017.
[49] P. Zhang, L. Wang, D. Wang, H. Lu, and C. Shen. Agile amulet: Real-time salient object detection with contextual attention. arXiv:1802.06960, 2018.

[50] R. Zhao, W. Ouyang, H. Li, and X. Wang. Saliency detection by multi-context deep learning. In CVPR, pages 1265–1274, 2015.

[51] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In ICCV, pages 2528–2535, 2013.