Context-Aware Attention Network for Salient Object Detection

Qinghua Ren, Shijian Lu, Jinxia Zhang, and Renjie Hu

***Abstract*—Benefiting from the discriminative feature extraction capability of convolutional neural networks (CNNs), deep learning techniques have recently achieved remarkable improvements in salient object detection. However, most of the existing deep models fall short in learning informative contextual features, leading to unsatisfactory results in challenging scenes. In addition, previous attention-guided saliency networks lack the ability to construct an effective bi-directional information propagation architecture. To address these issues, this paper proposes a context-aware attention network (CANet) that aims to build semantic connections between each pixel and its contexts. Specifically, we exploit two types of the information collection which can aggregate contextual features for each pixel and simultaneously transmit the semantic information of each position to other positions, thus resulting in a bi-directional structure. Besides, the proposed attention mechanism focuses on short- and long-range context regions to generate local and global attentive features. Last but not least, to maintain fine-grained spatial details, we design an attention-guided hierarchical network where the attended contextual information from deeper layers is transferred to shallower layers in a top-down manner. Extensive experiments on six popular saliency datasets show that our CANet performs favorably against the state-of-the-arts models in terms of various evaluation metrics.**

***Index Terms*—Deep learning, contextual information, visual attention, salient object detection.**

I. Introduction

SALIENT object detection (SOD), which aims to precisely segment foreground regions from backgrounds, has been becoming one of the fundamental challenges in computer vision. By modeling the human attention mechanism, SOD models can process an image into a probability map where the intensity value of each pixel represents the degree of how much the corresponding position draws human attention. By filtering out redundant background information and highlighting foreground objects, SOD has been proved to be a valuable pre-processing procedure in a wide range of object-related vision tasks, such as image retrieval [1], image quality assessment [2], image co-segmentation [3], weakly-supervised object detection [4], and object tracking [5], to name a few. Thus, to provide promising segmentation maps for these subsequent high-level applications, developing an accurate SOD model is of great significance.

Q. Ren is with the School of Electrical Engineering, Southeast University, Nanjing 210096, China (e-mail: renqinghua@seu.edu.cn).

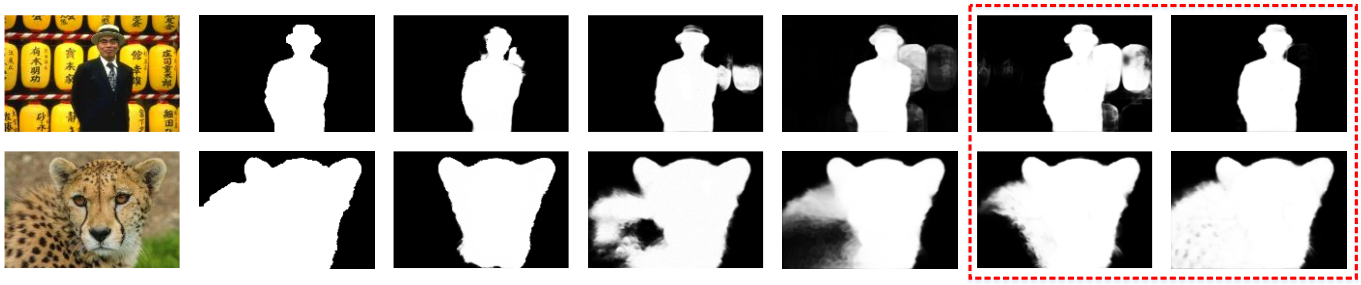
S. Lu is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: shijian.lu@ntu.edu.sg).

J. Zhang is with the School of Automation, Southeast University, Nanjing 210096, China (e-mail: jinxiazhang@seu.edu.cn).

R. Hu is with the School of Electrical Engineering, Southeast University, Nanjing 210096, China (e-mail: hurenjie@seu.edu.cn).

For a given image, conventional SOD approaches generally compute pixel-wise or region-wise saliency scores based on one or more low-level handcrafted features. The rationalities of these unsupervised methods are usually derived from various assumptions. For example, salient regions usually show high visual contrast with respect to their neighboring regions (local context [6]) and the entire image (global context [7]) in terms of some basic characteristics, e.g., color, texture, and intensity. Although such heuristic priors have been proven effective in extracting salient regions from relatively simple backgrounds, they are short of high-level semantic information, which may restrict the feature extraction ability to handle with complicated scenarios. It is crucial for SOD models to capture discriminative features instead of relying on data-driven statistics collected from some specific cases.

With the rapid advance of recent deep learning architectures, many significant progresses have been made in the computer vision community, such as instance classification, semantic segmentation, and salient object detection. Deep convolutional neural networks (CNNs), which have the powerful ability to automatically learn low-, middle-, and high-level features in a hierarchical structure, are capable of improving the detection performance by a large margin. Moreover, it is feasible and efficient to optimize the learnable parameters of a CNN model on large-scale image datasets. In terms of the fact that standard CNN models suffer from the damage of structural information in deeper layers, early SOD works based on fully convolutional networks (FCNs) [8] mainly concentrate on solving the blurry boundary problem by eliminating downsampling effects caused by multiple max-pooling operations. Hence, various refinement techniques were proposed, e.g., embedding over-segmentations [9], [10], [33], recurrent modules [11], [12], and hierarchical feature aggregation [13]–[18]. In detail, segment-wise labeling models generally compute the saliency score of each over-segmentation from local and global perspectives, but inevitably bringing about redundant computation. Apart from this, these previous approaches can hardly correct the errors caused by traditional image segmentation algorithms. Pixels in the same over-segmentation may be classified into different categories, though they share similar appearances in color space. The recurrent modules aim to eliminate inaccurate results generated by the previous block, but leading to a sharp increase in the running time cost. By contrast, hierarchical feature aggregation can make use of enriched semantic features from deeper layers and detailed boundary information from shallower layers, thus



(a) (b) (c) (d) (e) (f) (g)

Fig. 1. Visual comparison between the proposed CANet and some recent deep models. (a) Input images, (b) Ground truth, (c) DSS [13], (d) BMPM [22], (e) PiCANet [20], (f) Baseline hierarchical model, (g) Our attention network. We design the CANet by embedding three context-aware attention modules in the baseline hierarchical model. The saliency maps of these two models are highlighted in the red dashed bounding box. As can be seen, the CANet has better accuracy and completeness in complex scenarios.

leading to fine-grained saliency maps. Based on this knowledge, such refinement technique has raised a considerable amount of attention. More recently, researchers put more efforts on more elegant salient object detectors by incorporating attention mechanisms [19], [20], modified loss functions [15], [21], feature fusion strategy [16], [17], [22], [34], and global context information [12], [23]. Despite the superior performance achieved by the above-mentioned deep learning based works, it is still challenging for most of the existing deep models to alleviate distractions from cluttered backgrounds and highlight structurally complex objects, as shown in Fig. 1 (c)-(e).

In this paper, we propose a context-aware attention network (dubbed CANet) for saliency detection. The primary motivation behind this design is that the saliency of each pixel can be determined by its surroundings via information propagation. For each pixel, instead of simply assigning all pixels in its context region equal weights, we should properly enhance the effects of those semantically related pixels to help prediction. Meanwhile, each pixel should transmit its semantic information to other pixels for better inference. In short, the fundamental goal of our attention mechanism is to help each position connect with other positions by building an effective bi-directional information propagation architecture. Furthermore, we jointly incorporate both short- and long-range contextual information into the proposed attention module to self-adaptively learn more discriminative saliency features. Intuitively, short-range context concentrates on the local appearance, while long-range context analyzes global contrast differences. Similar multi-scale schemes have been widely used to further boost overall performance in many previous saliency works [10], [12], [18]. In the purpose of generating more robust feature descriptors, we also combine local dimension-reduced features with context-aware attention features, which shares the common principle with the residual learning framework [24]. Last but not least, to improve the sharpness of saliency results, we embed context-aware attention modules into the hierarchical network in a top-down fashion. In Fig. 1 (f), we enumerate some visual examples that our CANet is able to suppress noisy background responses and ensure good completeness of detected salient regions.

In summary, the contributions of this paper are three folds:
(1) The context-aware attention network is proposed to guide each pixel in building semantic connections with other pixels in local and global scopes. By learning complementary contextual features, the suggested attention model is able to filter out noisy background responses and uniformly highlight salient objects. Additionally, our attention modules enable joint training.
(2) The proposed CANet hierarchically aggregates the attended contextual features of deeper layers and low-level fine-grained features of shallower layers via short connections to sharpen the results, which significantly boosts SOD performance.
(3) Extensive experimental results on six popular benchmarks demonstrate that our CANet consistently outperforms the state-of-the-art deep learning based models.

II. Related Work

A comprehensive survey on traditional SOD algorithms can be found in [26]. Most recently, research on salient object detection is mainly dominated by CNN-based methods due to their strong feature extraction capability. In this section, we briefly review patch-wise deep SOD models, pixel-wise deep SOD models, and visual attention models.

A. Patch-wise Deep Models

Based on the success of CNNs in image classification [27], early patch-wise SOD models aim to remedy the deficiency of repeated downsampling operations in a generic CNN model. An image is divided into numerous over-segmentations (object proposals [9], or superpixels [10], [28], [29], [33]). As basic computational units, these over-segmentations are fed into the deep network one by one to compute saliency values. Wang *et al.* [9] jointly utilized local and global features by learning two individual deep networks to calculate the saliency score for each region. Li and Yu [10] extracted deep features for each superpixel in three different visual contexts, which can improve the performance to some extent. In [28], Zhao *et al.* designed a multi-context saliency network to construct multi-scale features based on local and global cues. Lee *et al.* [29] proposed a unified network to capture high-level CNN features and low-level handcrafted features. Although these former segment-wise labeling methods can outperform traditional SOD methods, they still fail to adequately incorporate spatial contexts and precisely locate salient objects in some challenging cases.

B. Pixel-wise Deep Models

To generate saliency maps in a direct way, pixel-wise deep SOD models replace original fully connected layer with fully convolutional layer. Most of these end-to-end models only

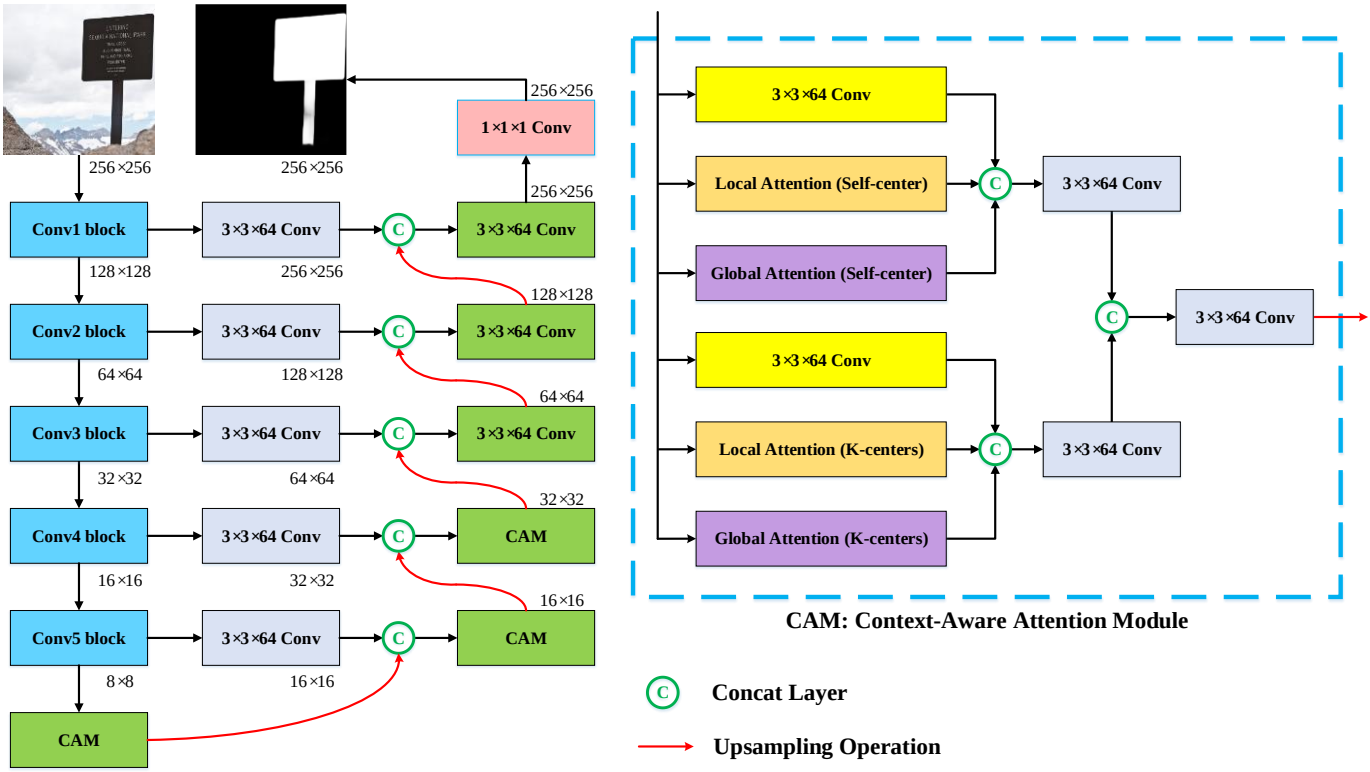


Fig. 2. Architecture of the proposed CANet. As shown in the blue dashed bounding box, we offer a detailed description of the context-aware attention module, which is embedded in three deeper layers of the decoder network. With the attention guidance, our hierarchical model can transfer more discriminative information to shallower layers, leading to more accurate saliency results.

require one feed-forward flow. They usually refine prediction maps by exploiting detailed boundary information in shallower stages. Liu and Han [13] developed a hierarchical refinement scheme to progressively sharpen saliency results. Li and Yu [14] designed a contrast-oriented saliency model which contains a multi-scale pixel-wise network and a segment-wise spatial pooling network. In [15], Luo *et al.* utilized a boundary loss to guide a non-local feature model in preserving the detailed structure of objects. Zhang *et al.* [16] modeled visual saliency by flexibly aggregating multi-level features at each resolution. Hou *et al.* [17] propagated high-level semantic information of deeper layers across all other shallower layers via dense short connections. More recently, Wang *et al.* [12] gathered multi-scale contextual information to better locate salient regions. In addition, they refined object boundaries by adopting a recurrent localization network and a local boundary refinement network. In [22], Zhang *et al.* extracted multi-context information using multiple dilated convolutions and designed a bi-directional structure to strengthen connections among different layers. In [23], Wang *et al.* used a pyramid pooling module to construct a multi-scale feature descriptor and a multi-stage refinement skill to obtain high-resolution prediction maps. Besides, some works focused on discovering new perspectives on pixel-wise SOD solutions. For example, Wang *et al.* [11] recurrently corrected prediction errors of the previous output until generating fine-grained saliency results in the last step. In [30], a deep unsupervised SOD method was proposed to optimize saliency network by learning with noisy unsupervised labels. Wang *et al.* [21] utilized fixation prediction to assist saliency estimation and introduced various metric-based loss functions for a significant performance boost.

C. Visual Attention Models

Recently, trainable attention models have been extensively explored in computer vision due to their effective and flexible mechanism. In [31], Kuen *et al.* progressively applied recurrent attention-based refinement on flexibly-sized image sub-regions. Chen *et al* [32] proposed the top-down reverse attention to learn more informative residual features. Zhang *et al* [19] designed an attention guided network which consists of spatial attention and channel-wise attention. In [20], Liu *et al* learnt attentive features to facilitate the final decision for each pixel. The global attention focused on larger contexts in deeper features while the local attention focused on smaller contexts in shallower features. Different from this previous work, we integrate local attention and global attention into a differentiable module instead of employing them separately, since local attention may be also beneficial for feature extraction in deeper layers. Aside from the research on salient object detection, deep attention models have also been studied in a wide range of vision tasks such as pose estimation [35], visual question answering [36], and image captioning [37], among others.

The proposed attention mechanism is partly motivated by the point-wise spatial attention network [25] for scene parsing. We further expand and improve it in the following three aspects. First, local contexts are integrated into the pixel-wise attention module in the purpose of exploiting richer local contextual features. The experiments are conducted in ablation studies to verify the effectiveness of the learned local attention. Second, we incorporate local attention and global attention into a unified

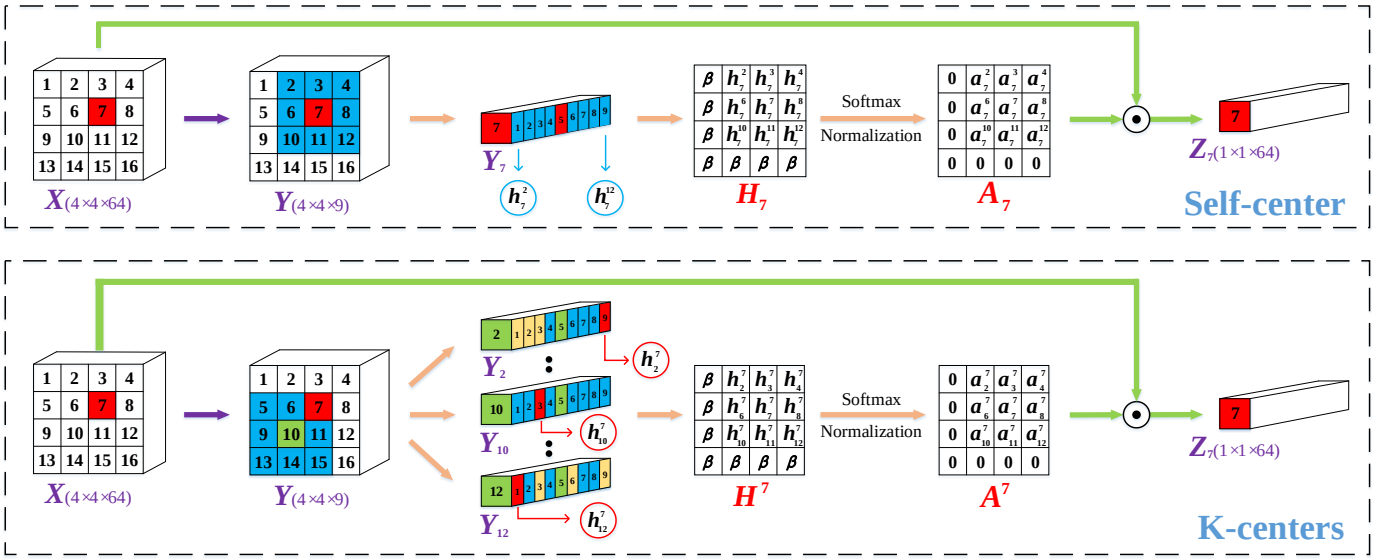


Fig. 3. Simplified illustration of the information collection. Based on the local attention, we explain the attention mechanism in the 'self-center' branch (top) and in the 'k-centers' branch (bottom). The difference of these two branches is the generation of attention weights.

module for more robust saliency inference. Last but not least, our context-aware attention modules are flexibly embedded into three deeper layers of the decoder network to produce fine-grained score maps with high accuracy.

III. The Proposed Model

A. Overall Architecture

Fig. 2 illustrates the overall architecture of the CANet, which consists of two components: context-aware attention module (CAM) and deep hierarchical network. The CAM is designed to extract more discriminative features by comprehensively exploiting the technique of information collection and the scope of context. More specifically, each CAM contains a 'self-center' branch and a 'k-centers' branch. These two parallel branches aim to help each position build semantic connections with its context regions via a bi-directional information propagation structure. From the perspective of the scale of context region, visual attention is subdivided into local attention and global attention, which can self-adaptively learn contextual features in a complementary way. Since the global attention generation is similar to the local attention generation, we will emphatically describe local attention in detail when giving an interpretation of the CAM. Our CAM is only employed in three deeper layers in term of superior performance and efficient computation.

The given image is uniformly resized to 256×256, and then fed into the attention-guided hierarchical network. By adopting a series of short connections, our hierarchical network is able to refine feature maps in a coarse-to-fine manner. Thus, through one feed-forward flow, the CANet can finally infer a pixel-level saliency map with the same resolution as the input image.

B. Context-aware Attention Module

Instead of treating all pixels in context regions equally, the proposed CAM aims at strengthening connections between each pixel and those semantically related pixels, thus resulting in more informative feature descriptors. Given convolutional features $\boldsymbol{X} \in \mathbb{R}^{W\times H\times C}$, where W, H and C indicate width, height and number of channels, respectively, our goal is to produce attentive features $\boldsymbol{Z} \in \mathbb{R}^{W\times H\times C}$ that emphasizes important foreground regions and suppresses noisy background responses. Considering that contextual information can play a significant role in saliency computation, we explore both short- and long-range context regions to generate local and global attention maps. In detail, for each position, we obtain a spatial attention map $\boldsymbol{A}$ with size $W\times H$, which assigns higher importance to the semantically related pixels. Subsequently, the attention map $\boldsymbol{A}$ is performed on features $\boldsymbol{X}$ to aggregate the information from other positions across all channels.

In this section, a simple example is firstly given to describe basic ideas of the 'self-center' branch and the 'k-centers' branch. Secondly, we make an elaborate explanation of the attention generation procedure. In the end, we introduce the common workflow of our context-aware attention module.

1) Simplified Description

To better understand the proposed attention mechanism, we assume the original feature representation to be $\boldsymbol{X} \in \mathbb{R}^{W\times H\times C}$ and focus on the $\hat{W}\times\hat{H}$ local neighboring region for information collection. Fig. 3 shows a simplification of the local attention generation at location i. In this example, we simply set $i = 7$, $W = 4$, $H = 4$, $C = 64$, $\hat{W} = 3$, $\hat{H} = 3$, and $C_{\text{local}} = 9$.

In the 'self-center' branch, a convolutional layer with 1×1 kernel is firstly applied to convert original features $\boldsymbol{X}$ into another features $\boldsymbol{Y} \in \mathbb{R}^{W\times H\times C_{\text{local}}}$ for channel adaption. According to the row-major order, we reshape the channels of $\boldsymbol{Y}_i$ to a spatial map with size $\hat{W}\times\hat{H}$. Therefore, each position in the $\hat{W}\times\hat{H}$ local neighboring region centered at location i has a corresponding channel. For example, the 2*nd* spatial position of $\boldsymbol{X}$ corresponds to the 1*st* channel of $\boldsymbol{Y}_7$, and the 12*th* spatial position of $\boldsymbol{X}$ refers to the 9*th* channel of $\boldsymbol{Y}_7$. Then we obtain a weight map $\boldsymbol{H}_i \in \mathbb{R}^{W\times H\times 1}$ where the index positions outside the $\hat{W}\times\hat{H}$ context region are set to a constant. The weight of the normalized attention map $\boldsymbol{A}_i \in \mathbb{R}^{W\times H\times 1}$ at location k is calculated

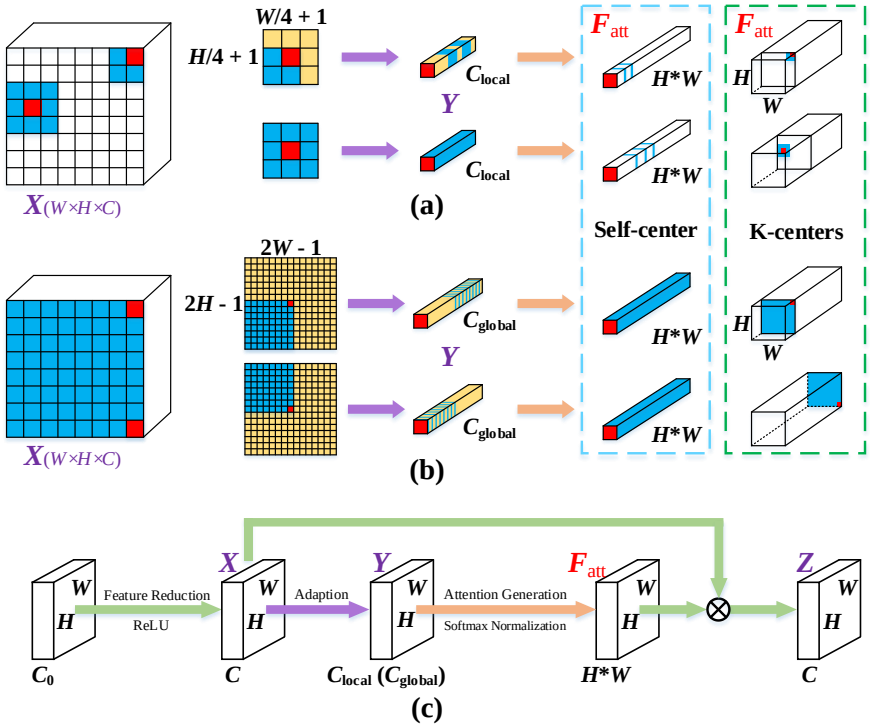


Fig. 4. Detailed structure of the CAM. (a) Local attention generation, (b) Global attention generation, (c) Common workflow. For two specific positions, we introduce the whole procedure of generating local/global attention weights Fatt in two branches. The attention weights for all positions can be computed in the same way. Finally, we can obtain new local/global attentive features.

via a softmax function:

$$a_i^k = \frac{\exp(h_i^k)}{\sum_{j=1}^{W*H} \exp(h_i^j)} \tag{1}$$

where $i, j, k \in \{1, 2, \ldots, W*H\}$ and $\boldsymbol{H}_i = \{ h_i^1, h_i^2, \ldots, h_i^{W*H} \}$. So we obtain a spatial attention map $\boldsymbol{A}_i = \{ a_i^1, a_i^2, \ldots, a_i^{W*H} \}$. The normalized attention weights for all locations can be calculated in the same way: $\boldsymbol{A} = \{\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_{W*H}\}$. Finally, the original features $\boldsymbol{X}$ in the channel c is weighted summed by $\boldsymbol{A}_i$ to generate the new features $\boldsymbol{Z}_i$:

$$Z_i(c) = \sum_{j=1}^{W*H} a_i^j \cdot X_j(c) \tag{2}$$

where $c \in \{1, 2, \ldots, C\}$. Similarly, we can get attentive features $\boldsymbol{Z} = \{\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_{W*H}\}$ that have the same size with $\boldsymbol{X}$.

In the 'k-centers' branch, our goal is to adopt a different way to generate attention weights for all pixels. We explore visual attention based on the knowledge that the information of each position can help the prediction of other positions [25]. By revisiting the relationship of the location i and its context, we observe that the 7*th* position is within the local $\hat{W} \times \hat{H}$ region centered at the 10*th* position, as illustrated in Fig. 3. More specifically, $\Omega(i)$ represents the context region centered at location i, and $\forall k \in \Omega(i)$ denotes all spatial positions in the neighboring region. Then we also follow the row-major order to determine the index number of i in $\Omega(k)$. For instance, the 10*th* position of $\boldsymbol{X}$ corresponds to the 3*rd* channel of $\boldsymbol{Y}_{10}$, and the 12*th* position of $\boldsymbol{X}$ refers to the 1*st* channel of $\boldsymbol{Y}_{12}$. Thus we obtain a weight map $\boldsymbol{H}_i$ and then compute the weight of the normalized attention map $\boldsymbol{A}_i$ at location k:

$$a_k^i = \frac{\exp(h_k^i)}{\sum_{j=1}^{W \times H} \exp(h_j^i)} \tag{3}$$

where $i, j, k \in \{1, 2, \ldots, W*H\}$. Finally, the attentive features $\boldsymbol{Z}_i$ in the channel c can be calculated by a dot product operation:

$$Z_i(c) = \sum_{j=1}^{W*H} a_j^i \cdot X_j(c) \tag{4}$$

where $c \in \{1, 2, \ldots, C\}$. In the same way, we obtain attended contextual features $\boldsymbol{Z} = \{\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_{W*H}\}$ in the 'k-centers' branch.

2) Attention Generation

In view of the scope of the context region Ω, visual attention can be subdivided into local attention and global attention. We empirically adopt the $(W/4+1)\times(H/4+1)$ neighboring region and the $(2W-1)\times(2H-1)$ over-completed region as local context and global context. In Fig. 4, it can be observed that the context region centered at some locations may cover some invalid areas (yellow areas). The channels of $\boldsymbol{Y}$, which are mapped to those invalid regions, will not participate in the back-propagation learning procedure. For the convenience of calculations, we reshape the normalized attention map $\boldsymbol{A}_i$ (or $\boldsymbol{A}^i$) with size $W \times H$ into the channels of $\boldsymbol{F}_{\text{att}} \in \mathbb{R}^{W \times H \times C_{\text{att}}}$ at location i, as illustrated in Fig. 4. More technically, the normalized weights in the 'self-center' branch are assigned to the corresponding channels of $\boldsymbol{F}_{\text{att}}$ at location i, while the normalized weights in the 'k-centers' branch are mapped to the corresponding spatial locations of $\boldsymbol{F}_{\text{att}}$ at channel i. Different from the global attention, numerous values of $\boldsymbol{F}_{\text{att}}$ in the local attention are constrained to be zero in the purpose of cutting off the information of those positions outside local neighboring context. To achieve this goal, we fill

in these corresponding channels with a large negative value β before softmax operation. In addition, the scalar values C_{local}, C_{global}, and C_{att} are constantly set to $(W/4+1)*(H/4+1)$, $(2W-1)*(2H-1)$, and $W*H$ respectively. The attentive features $\boldsymbol{Z} \in \mathbb{R}^{W\times H\times C}$ are obtained in one shot:

$$Z = X \otimes F_{\text{att}} \tag{5}$$

where $\otimes$ denotes a matrix multiplication in Caffe [38].

For original features $\boldsymbol{X}$, the proposed CAM can produce four types of attended contextual features: local attention in the 'self-center' branch, global attention in the 'self-center' branch, local attention in the 'k-centers' branch, and global attention in the 'k-centers' branch. When designing saliency attention networks, we comprehensively consider the range of context and the way of information collection. These context-aware features provide more discriminative saliency cues in a complementary way, which may be essential for a high-performance model.

3) Detailed Structure of the CAM

Fig. 4 (c) offers the common workflow of the local (global) attending operation which matches with local (global) attention block of CAM in Fig. 2 First, a convolutional layer with 3×3 kernels is employed to reduce the number of channels of the input features ($C_0 > C$). Then, based on the above-mentioned attention mechanism, we obtain spatial attention maps for all locations of the features $\boldsymbol{X}$. Finally, the normalized weights $\boldsymbol{F}_{\text{att}}$, which denote the relevance between each location and its contexts, operate on each spatial map of $\boldsymbol{X}$ across all channels to generate the new feature representation $\boldsymbol{Z}$.

Besides, another convolutional layer with 3×3 kernels is used for feature compression in each branch, as ablation experiments prove that these dimension-reduced features can contribute to better saliency estimation along with those contextual attention features. However, instead of making an intensive study on the optimal aggregation method, we simply incorporate the features in two parallel branches, as shown in Fig. 2. Afterwards, several concatenation operations and convolutional layers are used to aggregate all the features. Each convolutional layer is equipped with a rectified linear unit. As a result, the CAM can produce more informative feature descriptors with spatial size $W\times H$ and C channels. With our design, each position can build semantic connections with other positions, dramatically improving the accuracy of the deep saliency model. Essentially, the proposed CAM only has a series of convolutional layers, index operations, and softmax functions. Thus it is feasible and easy to train the whole attention network.

C. Attention-guided Hierarchical Network

Our saliency network is built upon the standard VGG-16 [27] backbone, which has been trained on the ImageNet dataset [39]. The encoder network of the CANet consists of 13 convolutional layers and 5 max-pooling layers from VGG-16. The resized input image with size 256×256 is fed into the network for feature extraction. Without the use of dilated convolutions [40], the spatial sizes of features derived from five conv modules are: 128×128, 64×64, 32×32, 16×16, and 8×8. It is well known that deeper layers can encode high-level semantic knowledge while shallower layers contain richer low-level structural information. Similar to many previous refinement techniques [13]–[18], we also use a pyramid-like structure to gradually sharpen feature maps in a coarse-to-fine manner. For the sake of efficient feature adaption, the last convolution layer of each conv module in the encoder part is followed by a 3×3 convolutional layer with 64 channels and a ReLU activation, thus resulting in the corresponding side-output feature. The spatial sizes of these side-output features are: 256×256, 128×128, 64×64, 32×32, and 16×16.

The decoder network of the CANet consists of six modules (see green blocks in Fig. 2), which generate feature maps with 64 channels, named Dec_6, Dec_5, Dec_4, Dec_3, Dec_2, and Dec_1. These feature maps are upsampled using bilinear interpolation by a factor of 2. Subsequently, we refine the enlarged feature map by concatenating it with the side-output feature in a step-by-step manner. Based on the knowledge that the powerful feature extraction in deeper layers is critical to the localization of foreground regions, we design an attention-guided model to transfer context-aware features from deeper layers to shallower layers in a top-down fashion. The proposed context-aware attention module is performed only in Dec_6, Dec_5, and Dec_4, while a 3×3 convolutional layer with 64 channels is adopted in Dec_3, Dec_2, and Dec_1. On the one hand, our CANet can gain the optimal performance in the current embedding setting, which will be fully investigated in ablation experiments. On the other hand, if the CAM is continued to apply in Dec_3 or more, the C_{global} in global attention will be very large which could sharply increase the computation cost.

In the training process, a one-channel convolutional layer with 1×1 kernel is performed on each decoding feature map. The deep supervision [41] is used to improve the convergence speed of the attention-guided hierarchical network. During the testing phase, an extra sigmoid activation function is employed on the Dec_1 to generate the final probability map.

IV. Experiments

A. Dataset and Setup

1) Datasets: Six public saliency datasets including ECSSD [42], HKU-IS [10], DUTS [43], DUT-O [6], PASCAL-S [44] and SOD [45] are selected to evaluate the proposed CANet. The ECSSD dataset consists of 1,000 natural images, most of which have semantically meaningful but structurally complex objects. HKU-IS contains 4,447 complex images, each of which has low visual contrast or multiple overlapping foreground objects. The DUTS dataset is one of the largest SOD benchmark datasets. It contains 10,553 training images (DUTS-TR), which usually are utilized for training deep learning based models. Its testing dataset (DUTS-TE), which includes 5,019 challenging images, is widely used to compare some high-performance models. The DUT-O is a representative SOD benchmark, which totally contains 5,168 complex images. This large dataset is valuable for performance comparison. The PASCAL-S dataset, which is collected from the PASCAL VOC segmentation dataset, has 850 challenging natural images. The last SOD dataset is built from the Berkeley Segmentation Dataset (BSD). This small dataset contains 300 images, most of which include cluttered

TABLE I
QUANTITATIVE RESULTS ON SIX DATASETS. THE TOP THREE MODELS ARE MARKED WITH **RED**, **GREEN** AND **BLUE**, RESPECTIVELY.

Method	Backbone	ECSSD [42]		HKU-IS [10]		DUTS-TE [43]		DUT-O [6]		PASCAL-S [44]		SOD [45]	
		maxF	MAE	maxF	MAE	maxF	MAE	maxF	MAE	maxF	MAE	maxF	MAE
DHS [13]	VGG16	0.907	0.059	0.891	0.052	0.812	0.065	-	-	0.830	0.096	0.823	0.127
DCL [14]	VGG16	0.901	0.068	0.893	0.063	0.786	0.082	0.757	0.080	0.816	0.116	0.831	0.131
NLDF [15]	VGG16	0.905	0.063	0.902	0.048	0.812	0.066	0.753	0.080	0.832	0.101	0.837	0.123
Amulet [16]	VGG16	0.915	0.059	0.896	0.052	0.778	0.085	0.743	0.098	0.839	0.099	0.803	0.141
DSS [17]	VGG16	0.921	0.052	0.911	0.040	0.825	0.057	0.781	0.063	0.840	0.098	0.843	0.122
SRM [23]	ResNet50	0.917	0.054	0.906	0.046	0.827	0.059	0.769	0.069	0.848	0.087	0.840	0.126
BMPM [22]	VGG16	0.928	0.045	0.921	0.039	0.852	0.049	0.774	0.064	0.863	0.074	0.852	0.106
DGRL [12]	ResNet50	0.922	0.041	0.910	0.036	0.828	0.050	0.774	0.062	0.856	0.072	0.843	0.103
PAGR [19]	VGG19	0.927	0.061	0.918	0.048	0.854	0.055	0.771	0.071	0.855	0.095	0.836	0.145
PiCANet [20]	VGG16	0.931	0.046	0.921	0.042	0.851	0.054	0.794	0.068	0.870	0.078	0.850	0.101
Ours	VGG16	0.938	0.044	0.930	0.037	0.876	0.044	0.810	0.058	0.880	0.075	0.865	0.099

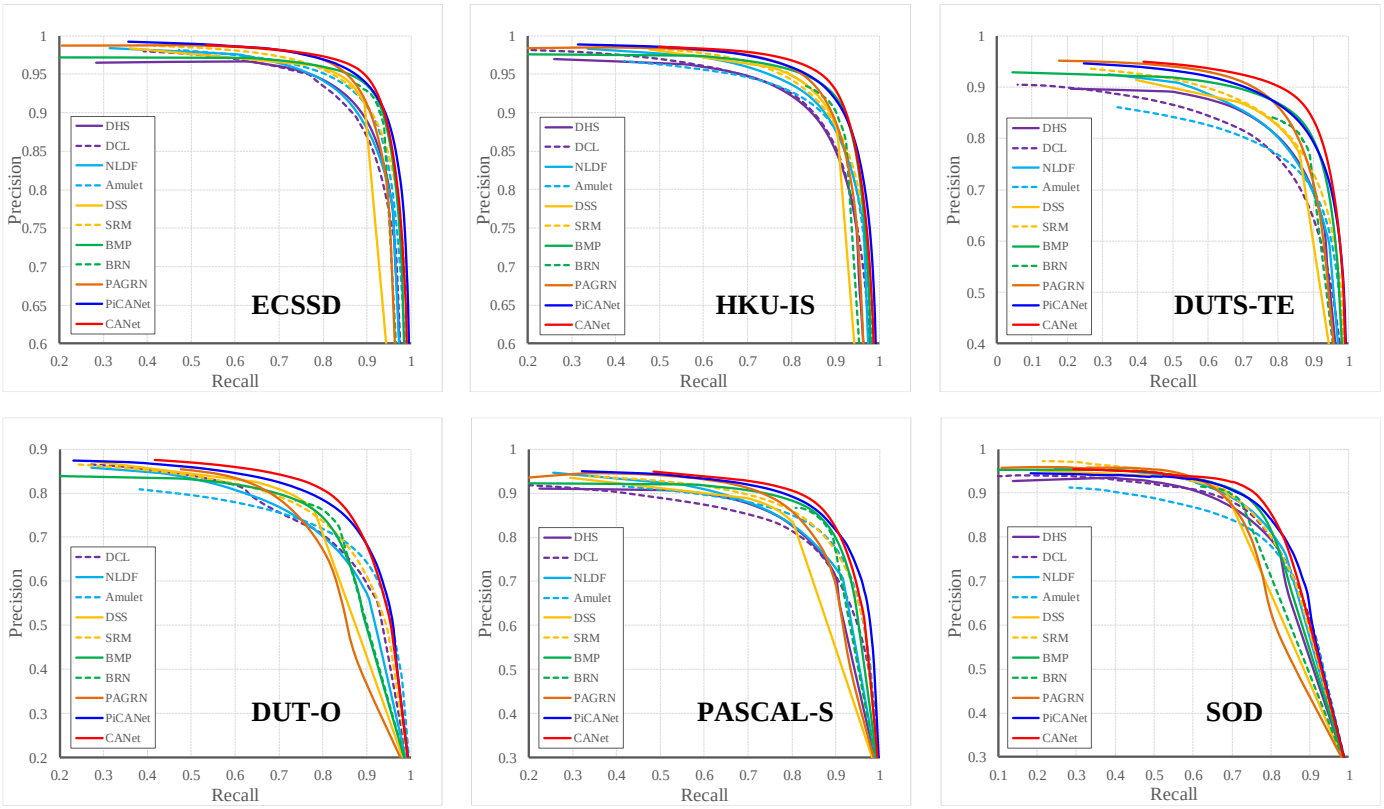


Fig. 5. Precision-recall curves on six benchmark datasets. Zoom-in for details.

backgrounds or complex salient objects. All the datasets offer the corresponding pixel-accurate ground truth annotations.

2) Evaluation Metrics: Three widely-used metrics, namely *PR-curve*, *F-measure*, and *MAE*, are adopted to evaluate the performance of our model and recent deep SOD models. Let $GT \in \{0,1\}$ and $SM \in (0,1)$ indicate the ground truth mask and the corresponding saliency map. By varying a fixed threshold from 0 to 255, a saliency map SM can be converted to some binary masks. A binary mask is denoted by $BM \in \{0,1\}$. The precision and recall are obtained by comparing BM and GT: $precision = |BM \cap GT|/|BM|$, and $recall = |BM \cap GT|/|GT|$, respectively. For a specific dataset, we compute a series of precision and recall pairs over all saliency maps, and then employ mean value pairs to draw the *PR curve*. The second metric *F-measure* score, which evaluates the comprehensive quality of saliency results, is defined as:

$$F_\beta = \frac{(1+\beta^2)\, precision \times recall}{\beta^2\, precision + recall} \tag{6}$$

where $\beta^2 = 0.3$ is used to emphasize precision more than recall as suggested in [46]. In this paper, we only report the *maximum F-measure* that can be computed from the *PR curve*. The third metric *MAE* (Mean Absolute Error) score is used to measure the dissimilarity between the binary ground truth mask *GT* and the predicted score map *SM* in a straight-forward way:

$$MAE = \frac{1}{W * H} \sum_{w=1}^{W} \sum_{h=1}^{H} \left| SM(w,h) - GT(w,h) \right| \tag{7}$$

where W and H indicate the width and height of *GT*. Similarly, we adopt a mean *MAE* score to evaluate the performance of one model on a dataset. Generally, a better model achieves higher *F-measure* score and lower *MAE* score.

3) Implementation Details: Our model is implemented on the Caffe [38] library. The training dataset of DUTS [43] is chosen to train our saliency network with standard binary cross entropy

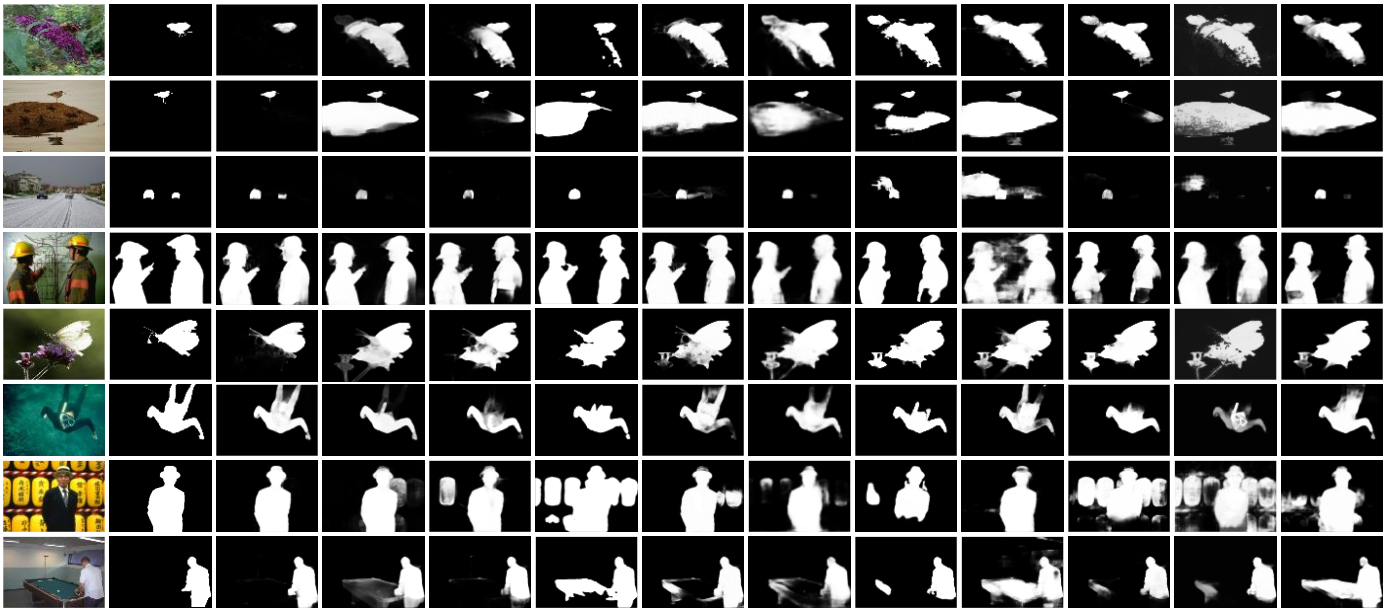


Fig. 6. Qualitative comparisons to some popular deep SOD models. Clearly, the proposed CANet can consistently produce more precise saliency results which are the closest to the ground truth annotations.

loss. Note that all the raw images and ground truth masks are uniformly resized to 256×256. The experiments are conducted using the Adam [47] optimizer with a weight decay of 1e-3. The parameters of the decoder network are randomly initialized and learned from scratch with an initial learning rate of 1e-5, while the parameters of the encoder network are initialized by the pre-trained VGG-16 [27] and fine-tuned with a 0.1 times smaller learning rate. Additionally, the learning rate is decreased by a factor of 10 after 10,000 steps. The maximum iteration step is set to be 20,000. It takes about 6 hours to train our CANet on a single GTX Titan X GPU. During the testing process, CANet achieves a real-time speed of 32 FPS, since it does not have any extra pre- or post-processing steps.

B. Comparison with the State of the Art

In this section, we compare the proposed CANet against 10 state-of-the-art deep models, including DHS [13], DCL [14], NLDF [15], Amulet [16], DSS [17], SRM [23], BMPM [22], DGRL [12], PAGR [19], and PiCANet [20]. To ensure a fair comparison against these existing methods, we use saliency results which are generated by the source code released by the authors. Note that non-deep learning models are not selected for performance comparison, since deep models outperform almost all of these classic algorithms by a large margin as reported in many previous works. Therefore, we prefer to compare our model with some leading CNN-based models.

1) Quantitative Performance Comparison: Table I and Fig. 5 show the quantitative comparison with other methods. In terms of the *maximum F-measure*, the proposed CANet achieves the best performance on all datasets and outperforms the second best model by 7.5‰, 9.8‰, 25.8‰, 20.2‰, 11.5‰, and 15.3‰ over ECSSD, HKU-IS, DUTS-TE, DUT-O, PASCAL-S, and SOD, respectively. In particular, the superior performance of the CANet on two large-scale saliency datasets (DUTS-TE and DUT-O) demonstrates that our attention-guided model has more discriminative ability to segment salient regions in complex scenes. Comparing *MAE* scores, our method ranks first on DUTS-TE, DUT-O, and SOD. Although the *MAE* score of the recent model DGRL [12] is a bit lower than ours on ECSSD, HKU-IS, and PASCAL-S, our saliency model greatly outperforms it with respect to the *maxF* score and *PR curve*. Fig. 5 shows PR curves on six benchmark datasets. It can be easily seen that the CANet performs favorably against other existing models on all datasets.

2) Qualitative Performance Comparison: In Fig. 6, we give the qualitative results for more intuitive comparison. It can be clearly found that our CANet can uniformly highlight salient objects and filter out the redundant background information in various complex scenes, such as images with one or more small objects (row 1, 2, and 3), objects touching the image boundary (row 4, 6, 7, and 8), cluttered backgrounds (row 1, 5, and 7), multiple large objects (row 4), low color contrast (row 3 and 6), non-venter bias (row 8). From these visual examples, we have the following two observations. The one is that some models often suffer from the distractions of complex backgrounds, thus falsely assigning higher saliency values to a part of non-salient regions. The other one is that a portion of foreground regions fail to be highlighted by some methods. By contrast, our model achieves the best performance with the guidance of the context-aware attention mechanism that is able to transmit the semantic information of each position to other positions. It is worth noting that the deep hierarchical network plays a significant role to improve the fine-grained spatial structures of objects.

C. Ablation Studies

In this section, we mainly evaluate the contribution of each component in the proposed context-aware attention module. All experiments are performed on two large-scale saliency datasets DUT-O [6] and DUTS-TE [43]. We adopt two common metrics *maximum F-measure* and *MAE* to evaluate the effectiveness of

TABLE II
ABLATION ANALYSIS OF DIFFERENT DESIGN OPTIONS IN CAM. THE FINAL CONFIGURATION IS MARKED WITH **RED**.

No.	Self-center	K-centers	Feature Compression	Local Attention	Global Attention	DUT-O [6]		DUTS-TE [43]	
						MaxF	MAE	MaxF	MAE
1	✓	✓	✓	×	×	0.782	0.063	0.854	0.049
2	✓	✓	×	✓	×	0.797	0.059	0.866	0.047
3	✓	✓	×	×	✓	0.801	0.060	0.866	0.047
4	✓	✓	✓	✓	×	0.801	0.058	0.870	0.045
5	✓	✓	✓	×	✓	0.806	0.059	0.869	0.045
6	✓	✓	×	✓	✓	0.801	0.062	0.868	0.048
7	✓	✓	✓	✓	✓	**0.810**	**0.058**	**0.876**	**0.044**
8	✓	×	✓	✓	✓	0.804	0.058	0.870	0.045
9	×	✓	✓	✓	✓	0.805	0.059	0.870	0.046

TABLE III
ABLATION STUDY OF EMBEDDING CHOICES. THE BEST CHOICE IS MARKED WITH **RED**.

Module	Dec_6	Dec_5	Dec_4	DUT-O [6]		DUTS-TE [43]	
Resolution	8×8	16×16	32×32				
C_{local}	9	25	81	MaxF	MAE	MaxF	MAE
C_{global}	225	961	3969				
No.1	✓	×	×	0.804	0.059	0.870	0.045
No.2	✓	✓	×	0.809	0.058	0.872	0.045
No.3	✓	✓	✓	**0.810**	**0.058**	**0.876**	**0.044**

different design options. The quantitative evaluation results are summarized in Table 2 and Table 3. Note that all variants are designed based on the hierarchical network. We do not conduct more experiments to show the effect of the hierarchical network and the deep supervision training strategy, as they have been proven effective in many previous works.

1) Local Attention: To validate the effectiveness of the local attention, we replace the CAM with other modified versions, as shown in Table 2. The No.1 setting refers to a model without any attention scheme. The No.2 setting corresponds to a model only with the local attention. By combing the local dimension-reduced features and the local attentive features, we can obtain a modified module (No.4 setting). It can be clearly seen that the No.4 setting consistently achieves best performance, compare to No.1 and No.2. This indicates that our local attention is useful to extract local information for better saliency inference. Note that these three settings do not consider the global attention. To investigate whether local attention is also complementary to the global attention, we also add the global attention into No.1 and No.4 to get two new versions No.5 and No.7. The comparison results further confirm that local attention can boost the overall performance. In terms of *maxF* score and *MAE* score, we can easily obverse that No.5 has already outperforms state-of-the-art deep models by a large margin, as shown in Table 1. It is well known that improving an existing high-performance model is not easy. Despite that, our final setting No.7 can make further improvement on No.5 (DUT-O: 0.806 → 0.810, and DUTS-TE: 0.869 → 0.876) in terms of *maxF* score. This demonstrates that the proposed local attention contributes to saliency estimation.

2) Global Attention: Similar to various design options in the local attention, we also study the effect of global attention in two aspects: *w/o* local attention and *w/* local attention. Table 2 provides detailed evaluation results of different configurations. Compared with No.1 and No.3, the No.5 variant has the best performance on both two large-scale datasets. As can be seen, the combination of the global attention and local dimension-reduced features can achieve the highest *maxF* score on DUT-O when we randomly combine any two components (No.4, No.5, and No.6). The underlying reason is that the largest receptive field from global attention and the smallest receptive field from local features may gain a maximum cooperation effect. Furthermore, we also see a large increase in performance when we add global attention into the variant No.4. This proves that global attention can effectively promote the information propagation over the entire image, which is crucial for powerful feature extraction.

3) Information Collection: To further study the influence of two branches in the CAM for salient object detection, we retrain another two modified models with the same training dataset and setting. The No.8 represents a model only with the 'self-center' branch, and the No.9 refers to a model only with the 'k-centers' branch. Our final setting in the CAM corresponds to the No.7 which contains both two parallel branches. The quantitative comparison results of these three variants are depicted in Table 2. Clearly, the No.7 performs favorably against the other two ones. This confirms that exploring different ways of collecting information is beneficial for bi-directional message passing.

4) Embedding Choices: The baseline model refers to the No.1 setting in Table 2. Based on the baseline model, we try to embed the CAM in the decoder network with a top-down sequence. As mentioned above, C_{local} and C_{global} in the CAM are decided by the resolution of input features. More details can be found in Table 3. The C_{global} value will be set to be 16,139 if we continue to apply the CAM in Dec_3. That may greatly increase the computation cost. Thus, to achieve a better tradeoff between performance and computation, we only compare three different

embedding choices. Even though only Dec_6 is equipped with the CAM, we can obverse a great improvement on the baseline model. As shown in Table 3, the No.3 is the best embedding choice for our model in terms of *maxF* score and *MAE* score.

V. Conclusion

In this paper, we have developed a context-aware attention network for salient object detection. By exploring the way of information collection and the scope of context, our CANet can effectively enhance the ability to identify salient objects, thus generating accurate prediction results. In detail, with the guidance of the proposed spatial attention mechanism, we can selectively aggregate contextual information for each pixel, which is able to filter out redundant background information and ensure good completeness. By learning attentive contextual features in a complementary way, more discriminative saliency cues can be encoded into three deeper layers of the decoder network for a clear performance boost. Moreover, the CANet combines high-level semantic knowledge from deeper layers and low-level boundary information from shallower layers via a series of short connections, which can progressively sharpen saliency results. Exhaustive experiments with 10 state-of-the-art methods on six saliency datasets demonstrate the superior performance of our CANet. In the future, we intend to improve SOD performance by designing better loss functions and more sophisticated saliency networks.

References

[1] Y. Gao, M. Shi, D. Tao, and C. Xu, "Database saliency for fast image retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 359–369, Mar. 2015.

[2] W. Zhang, A. Borji, Z. Wang, P. Le Callet, and H. Liu, "The application of visual saliency models in objective image quality assessment: A statistical evaluation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1266–1278, Jun. 2016.

[3] K. Jerripothula, J. Cai, and J. Yuan, "Image co-segmentation via saliency co-fusion," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1896–1909, Sep. 2016.

[4] Y. Tang *et al.*, "Weakly supervised learning of deformable part-based models for object detection via region proposals," *IEEE Trans. Multimedia*, vol. 19, no. 2, pp. 393–407, Feb. 2017.

[5] C. Ma, Z. Miao, X.-P. Zhang, and M. Li, "A saliency prior context model for real-time object tracking," *IEEE Trans. Multimedia*, vol. 198, no. 11, pp. 2415–2424, Nov. 2017.

[6] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3166–3173.

[7] M. Cheng, N. Mitra, X. Huang, P. Torr, and S. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[9] L. Wang, H. Lu, X. Ruan, and M. H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3183–3192.

[10] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5455–5463.

[11] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 825–841.

[12] T. Wang, L. Zhang, S.Wang, H. Lu, G. Yang, X. Ruan, and A. Borji, "Detect globally, refine locally: A novel approach to saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp 3127–3135.

[13] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 678–686.

[14] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 478–487.

[15] Z. Luo *et al.*, "Non-local deep features for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6593–6601.

[16] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 202–211.

[17] Q. Hou *et al.*, "Deeply supervised salient object detection with short connections," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5300–5309.

[18] Q. Ren and R. Hu, "Multi-scale deep encoder-decoder network for salient object detection," *Neurocomputing*, vol. 316, pp. 95-104, Nov. 2018.

[19] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 714–722.

[20] N. Liu, J. Han, and M.-H. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3089–3098.

[21] W. Wang, J. Shen, X. Dong, and A. Borji, "Salient object detection driven by fixation prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1171–1720.

[22] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang, "A bi-directional message passing model for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1741–1750.

[23] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, "A stagewise refinement model for detecting salient objects in images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4039–4048.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[25] H. Zhao *et al.*, "PSANet: Point-wise Spatial Attention Network for Scene Parsing," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 270–286.

[26] A. Borji, M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learning Represent.*, 2014, pp. 1–14.

[28] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1265–1274.

[29] G. Lee, Y. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 660–668.

[30] J. Zhang *et al.*, "Deep unsupervised saliency detection: A multiple noisy labeling perspective," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9029–9038.

[31] J. Kuen, Z. Wang, and G. Wang, "Recurrent attentional networks for saliency detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3668–3677.

[32] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 234–250.

[33] K. Fu, Q. Zhao, and I. Y.-H. Gu, "Refinet: A deep segmentation assisted refinement network for salient object detection," *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 457–469, Feb. 2019.

[34] H. Xiao, J. Feng, Y. Wei, M. Zhang, and S. Yan, "Deep salient object detection with dense connections and distraction diagnosis," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3239–3251, Dec. 2018.

[35] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1831–1840.

[36] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 21–29.

[37] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5659–5667.

[38] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Multimedia*, 2014, pp. 675–678.

[39] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
[40] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2018.
[41] C.-Y. Lee, S. Xie, P. W. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2015, pp. 562–570.
[42] J. Shi, Q. Yan, L. Xu, and J. Jia, "Hierarchical image saliency detection on extended CSSD," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 4, pp. 717–729, Apr. 2016.
[43] L.Wang *et al.*, "Learning to detect salient objects with image-level supervision," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 136–145.
[44] Y. Li, X. Hou, C. Koch, J. Rehg, and A. Yuille, "The secrets of salient object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 280–287.
[45] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2001, pp. 416–423.
[46] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1597–1604.
[47] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learning Represent.*, 2015.