This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2997178, IEEE Transactions on Multimedia

Salient Object Detection by Fusing Local and Global Contexts

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

***Abstract*—Benefiting from the powerful discriminative feature learning capability of convolutional neural networks (CNNs), deep learning techniques have achieved remarkable performance improvement for the task of salient object detection (SOD) in recent years. However, most existing deep SOD models do not fully exploit informative contextual features, which often leads to suboptimal detection performance in the presence of a cluttered background. This paper presents a context-aware attention module that detects salient objects by simultaneously constructing connections between each image pixel and its local and global contextual pixels. Specifically, each pixel and its neighbors bidirectionally exchange semantic information by computing their correlation coefficients, and this process aggregates contextual attention features both locally and globally. In addition, an attention-guided hierarchical network architecture is designed to capture fine-grained spatial details by transmitting contextual information from deeper to shallower network layers in a top-down manner. Extensive experiments on six public SOD datasets show that our proposed model demonstrates superior SOD performance against most of the current state-of-the-art models under different evaluation metrics.**

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

I. Introduction

SALIENT object detection (SOD) seeks the precise segmentation of foreground objects from the image background, which has become one of the major challenges in computer vision research in recent years. Inspired by the human attention mechanism, SOD models convert an image into a probability map that describes how much each pixel draws our attention. By filtering out the redundant background information, SOD has been proven to be a powerful preprocessing component in a wide range of vision tasks such as image retrieval [1], image quality assessment [2], image co-segmentation [3], weakly supervised object detection [4], and object tracking [5]. Thus, developing accurate and robust SOD models is of great importance to these types of high-level tasks.

This work was supported in part by the scholarship from China Scholarship Council under the Grant No. 201906090194, in part by the NTU Start-up Grant M4082034, in part by the National Natural Science Fund of China No. 61703100, in part by the Natural Science Foundation of Jiangsu No.BK20170692, in part by the Fundamental Research Funds for the Central Universities, and in part by the Big Data Computing Center of Southeast University.

Q. Ren and R. Hu are with the School of Electrical Engineering, Southeast University, Nanjing 210096, China (e-mail: renqinghua@seu.edu.cn; hurenjie@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: jinxiazhang99@163.com).

Given an image, conventional SOD models [26] usually compute the pixel-wise or region-wise saliency based on one or more handcrafted low-level features. The rationale is based on different assumptions; e.g., salient regions usually show high visual contrast with respect to their neighboring regions (the local context [6]) or the entire image (the global context [7]) in terms of certain low-level features such as the color, texture, and intensity. Although these types of heuristic priors have been proven effective at extracting salient regions from simple image backgrounds, they lack high-level semantic information, which often restricts the feature extraction capability in dealing with images with complicated backgrounds. It is crucial for SOD to capture the discriminative features instead of just relying on handcrafted low-level features.

Various deep convolutional neural networks (CNNs) have been developed for the SOD task in recent years. The CNN-based SOD is capable of learning robust features at different hierarchical levels, and it has improved the SOD performance by a large margin relative to the handcrafted features. Since CNNs tend to discard structural details in deeper layers, early CNN-based SOD focused more on solving the blurry boundary problem by alleviating the down-sampling effects that are introduced by multiple max-pooling operations. Different refinement techniques have been proposed, including embedding over-segmentations [9], [10], [33], recurrent modules [11], [12], hierarchical feature aggregation [13]-[18], etc. The recent CNN-based SOD research focuses more on accurate SOD by incorporating attention mechanisms [19], [20], modified loss functions [15], [21], feature fusion strategies [16], [17], [22], [34], global context information [12], [23], etc. Although these approaches have achieved impressive results, they still face various problems in dealing with highly cluttered image backgrounds or foreground objects with complicated structures, as illustrated in Figs. 1 (c)-(e).

In this paper, we propose the Context-aware Attention Network (CANet) for accurate salient object detection in the presence of cluttered backgrounds and complicated object structures. The major idea is that the saliency of an image can be better predicted by bidirectionally propagating the semantic information between each pixel and its contextual pixels. Furthermore, the contextual pixels of each pixel should not be treated equally; instead, algorithms should favor those pixels that are semantically related throughout the information propagation process. In addition, CANet incorporates both short-range and long-range contextual information to concurrently capture the local appearance and global contrast, which helps to learn more discriminative saliency features effectively. Finally, we design an attention-guided hierarchical

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2997178, IEEE Transactions on Multimedia

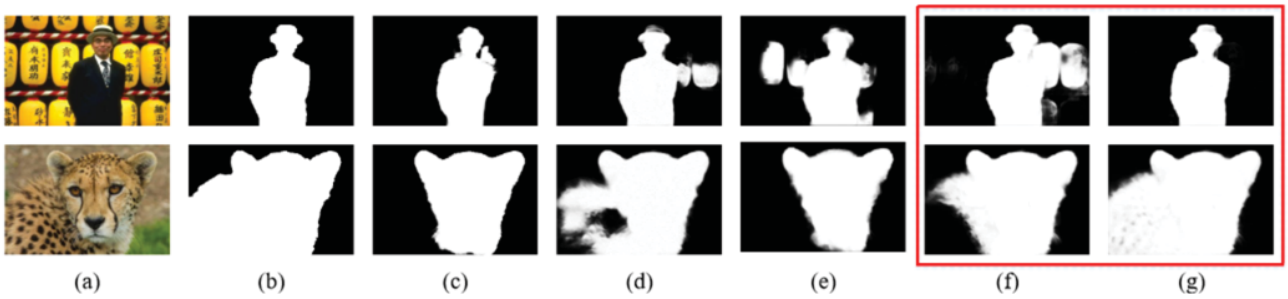

Fig. 1. The proposed CANet produces more accurate predictions of salient objects: for the input images in (a) and the corresponding ground-truth annotations in (b), (c)-(g) show the prediction of salient objects by DSS [17], BMPM [22], PAGENet [49], a baseline hierarchical model, and our CANet, respectively. In (g), CANet incorporates three context-aware attention modules over the baseline hierarchical model in (f), as highlighted in the red bounding box.

network structure that propagates contextual information from deeper layers to shallower layers in a top-down manner. This hierarchical design helps to greatly suppress the prediction uncertainty in the presence of highly cluttered backgrounds or foreground objects with complex structures, as illustrated in Fig. 1.

The contributions of this work are threefold. First, we design a context-aware attention module that guides each pixel to build connections with its contextual pixels in two network branches. Second, both local and global contexts are exploited to learn complementary contextual features that help filter out responses from noisy backgrounds and highlight salient foreground objects effectively. Third, extensive experiments on six public datasets show that CANet achieves superior SOD performance relative to the current state-of-the-art algorithms.

The rest of this paper is organized as follows. Section II gives a short overview of the related works on deep SOD networks and visual attention models. In Section III, the proposed context-aware attention module and deep hierarchical network are presented in detail. Section IV presents the experimental results, including a performance comparison with the state-of-the-art algorithms and ablation studies. Finally, a few concluding remarks are presented in Section V.

II. Related Work

Most recent SOD works make use of CNNs due to their strong feature extraction capability. This section will briefly review the recent SOD research, including patch-wise deep SOD models, pixel-wise deep SOD models, and visual-attention-based SOD models, respectively.

A. Patch-wise Deep Models

With the success of CNNs in image classification [27], early patch-wise SOD research intended to compensate for the effects of repeated down-sampling operations such as those performed in generic CNNs. The typical approach is to divide an image into numerous segments such as object proposals [9] and superpixels [10], [28], [29], [33]. As basic computational units, these divided segments are fed individually into deep networks to compute the saliency. For example, [9] jointly utilizes local and global features by learning two individual deep networks to calculate the saliency score for each region. [10] extracts deep features for each superpixel in three different visual contexts. [28] designs a multi-context saliency network to construct multi-scale features based on local and global cues. In addition, [29] proposes a unified network to capture high-level CNN features and low-level handcrafted features.

Although these segment-wise labeling models outperform traditional SOD methods, they do not incorporate sufficient spatial contexts and often fail to locate salient objects precisely in challenging images. Moreover, these models are computationally intensive. To address these issues, most later deep models adopt the hierarchical feature aggregation based on fully convolutional networks (FCNs) [8].

B. Pixel-wise Deep Models

Pixel-wise SOD models replace fully connected layers with fully convolutional layers to predict the saliency in a direct way. These models only require one feed-forward flow, and they usually refine their predictions by exploiting detailed boundary information in shallower stages. For example, [13] develops a hierarchical refinement scheme to progressively sharpen the saliency predictions. [14] designs a contrast-oriented saliency model that contains a multi-scale pixel-wise network and a segment-wise spatial pooling network. [15] utilizes a boundary loss to guide a non-local feature model to preserve detailed object structures. Furthermore, [16] models the visual saliency by flexibly aggregating multi-level features at each resolution, and [17] propagates the high-level semantic information of deeper layers across all of the other shallower layers via dense short connections. More recently, [12] gathers multi-scale contextual information to better locate salient regions. [22] extracts contextual information using multiple dilated convolutions and designs a bi-directional structure to strengthen the connections among different layers. [23] uses a pyramid pooling module to construct a multi-scale feature descriptor and develop a multi-stage refinement skill to obtain high-resolution predictions. In addition, some works have focused on discovering new perspectives on pixel-wise SOD solutions. For example, [11] recurrently corrects the prediction errors of the previous output until fine-grained predictions are generated in the last step. [30] optimizes the saliency network by learning with noisy unsupervised labels. Moreover, [21] utilizes fixation prediction [35], [36] to assist the saliency estimation and introduces various metric-based loss functions to achieve better saliency prediction. [51]

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2997178, IEEE Transactions on Multimedia

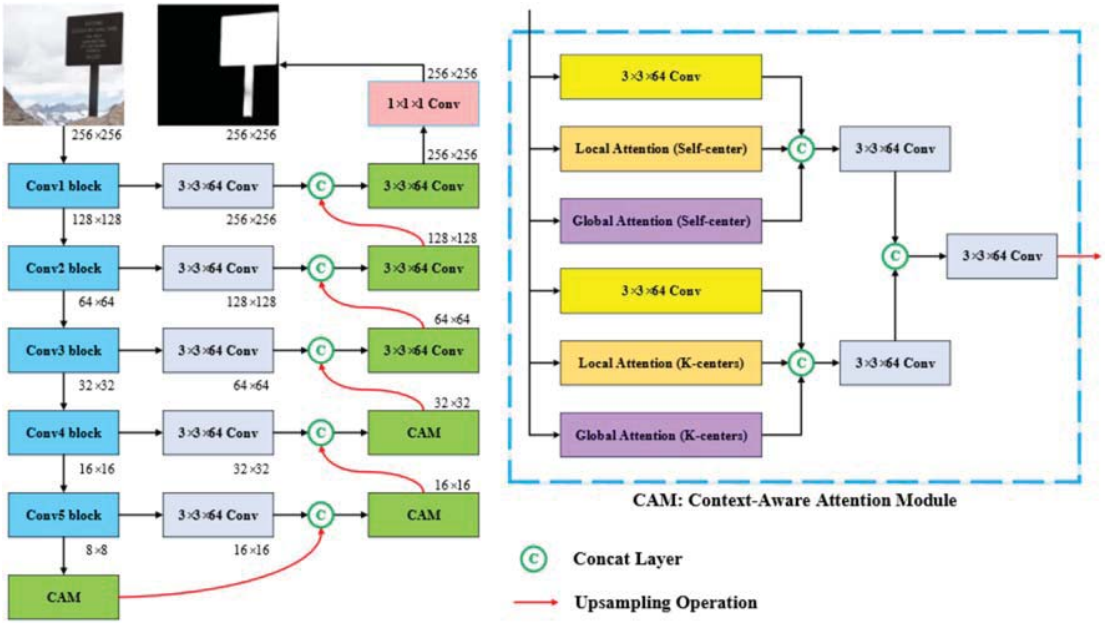

Fig. 2. Architecture of the proposed CANet: CANet incorporates three context-aware attention modules in both the 'self-center' branch and the 'k-centers' branch (the details are shown in the dashed bounding box on the right) in three deep layers of the decoder network. Using the context-aware attention module, CANet can transfer more discriminative information to shallower layers hierarchically and produce more accurate predictions of salient objects.

improves the boundary quality by designing a novel hybrid loss. [52] guides the deep network to learn the saliency by incorporating edge features.

Although the above SOD models have achieved very impressive saliency predictions, they do not sufficiently exploit the informative contextual features in deeper layers. In our proposed CANet, a powerful feature extraction module is designed that explicitly collects useful contextual information in deeper stages and greatly improves the saliency prediction.

C. Visual Attention Models

Visual attention models have been explored extensively for the SOD task in recent years. For example, [31] progressively applies recurrent attention-based refinements on flexibly-sized image sub-regions. [32] proposes top-down reverse attention to learn more informative residual features. [19] designs an attention-guided network that consists of spatial attention and channel-wise attention. [48] develops a pyramid feature attention network to generate the contextual saliency features, and [49] exploits multi-scale attentive features with an enlarged receptive field. Moreover, [20] learns the attentive features to facilitate the final decision for each pixel. The major idea is to learn the global attention to capture the large-scale contexts in deeper layers and to learn the local attention to capture the neighboring small-scale contexts in shallower layers. This type of attention-based SOD model learns the global and local attentions in deeper and shallower layers separately, but it neglects the fact that deeper layers may also learn local attention and vise versa. In addition to the SOD task, deep attention models have also been studied in many other vision tasks such as pose estimation [56], visual question answering [57], and image captioning [37].

Most existing attention models compute a single spatial attention map and obtain the weighted features, and the relations among the neighboring pixels of different distances are largely neglected. Additionally, many existing models capture the global attention while ignoring the effect of the local attention. To address these problems, we explicitly develop a customized attention map for each individual pixel to strengthen the connections among the pixels that belong to the same category. We also integrate the local attention and global attention into a differentiable module that greatly improves the inference in saliency modeling.

III. The Proposed Model

A. Overall Architecture

Fig. 2 shows the overall architecture of our proposed CANet, which consists of two major components: a context-aware attention module (CAM) and a deep hierarchical network. The CAM is designed to learn the discriminative saliency features by exploiting the contextual information of different spatial spreads comprehensively. Specifically, each CAM consists of a 'self-center' branch and a 'k-centers' branch in parallel that collaborate to learn the connections between each pixel and the contextual pixels for their bidirectional information propagation. Within each branch, local attention and global attention are employed to capture the complementary contextual features on different scales adaptively. Since the global attention and local attention are implemented in similar network structures, we will present the proposed CAM by giving a detailed description of the local attention in the ensuing subsection. Note that this CAM is only employed in three deeper layers to achieve both superior performance and efficient computation.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2997178, IEEE Transactions on Multimedia

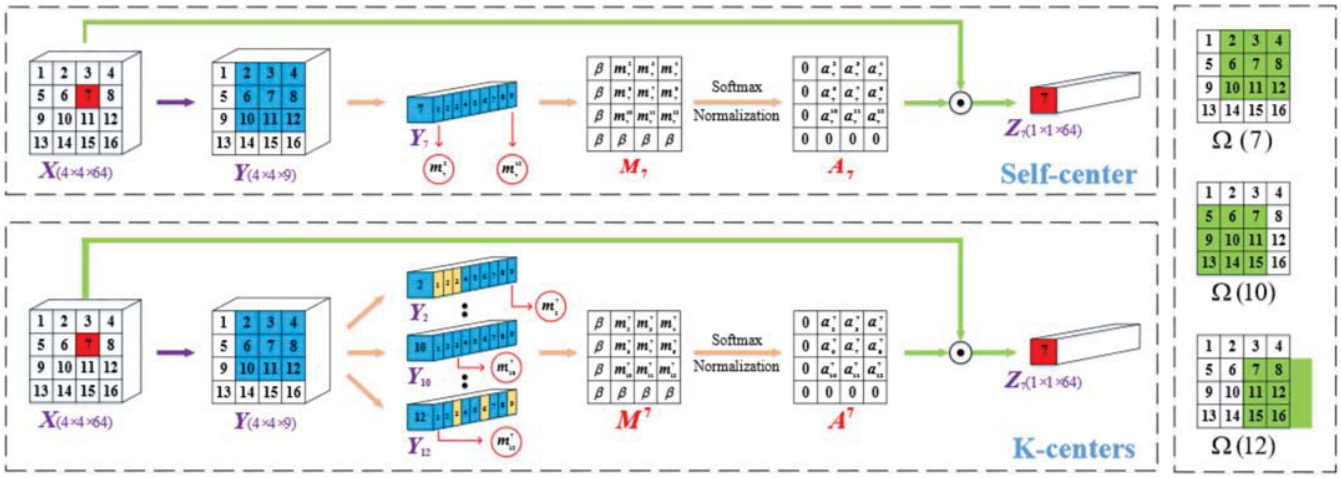

Fig. 3. Simple illustration of the attention generation for a target pixel: the context-aware attention is learned by a 'self-center' branch on the top and a 'k-centers' branch on the bottom in parallel (the high-level structure is shown in the dashed box in Fig. 2). The differences between these two branches produce attention weights. Here, $\Omega(i)$ denotes the 3×3 local context region at pixel location i. Three illustrations of these local regions are provided on the right.

As shown in Fig. 2, each training image is uniformly resized to 256×256 and fed into the attention-guided hierarchical network for training. By adopting a series of short connections, our hierarchical network is capable of refining feature maps in a coarse-to-fine manner. Therefore, CANet can infer a pixel-level saliency map with the same resolution as the input image via a single feed-forward flow.

B. Context-aware Attention Module

Instead of treating all of the pixels in the context regions equally, the proposed CAM is designed to strengthen the connections between each pixel and the semantically related pixels to produce more informative feature descriptors. Given convolutional features $X \in \mathbb{R}^{W \times H \times C}$, where W, H and C indicate the width, height and number of channels, respectively, our goal is to produce attentive features that emphasize the important foreground regions and suppress noisy background responses. Because the contextual information can play a significant role in the saliency computation, we explore both short-range and long-range context regions to generate local and global attention maps. For each position, we obtain a spatial attention map A of size $W \times H$ that assigns higher importance to the semantically related pixels. Subsequently, the attention map A is performed on features X to aggregate the information from the other positions across all of the channels. Note that we explore two independent ways to compute the attention weights of each pixel through a 'self-center' branch and a 'k-centers' branch, as illustrated in Fig. 3. The attentive features generated from these two branches are complementary to each other and improve the overall SOD performance cooperatively.

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

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

In the 'self-center' branch, a convolutional layer with a 1×1 kernel is initially applied to convert the original features X into other features $Y \in \mathbb{R}^{W \times H \times C_{\text{local}}}$ for the channel adaption. Here, Y_i denotes all of the channels of pixel i in Y. $\Omega(i)$ represents the $\overline{W} \times \overline{H}$ context region centered at pixel location i, and $\forall k \in \Omega(i)$ denotes all of its contextual pixels. Three illustrations are provided on the right of Fig. 3. For simplification, when the $7th$ position is regarded as the target pixel (the red node in X), $\forall k \in \Omega(7)$ (green nodes) indicates all of this pixel's contextual pixels (the blue nodes in Y). The weight for each contextual pixel is derived from Y_7 according to a row-major order. For example, the weight of spatial position 2 in X is the first channel of Y_7, and the weight of spatial position 12 in X is the last channel of Y_7. More generally, Y_i denotes the correlation of target pixel i and its contextual pixels. The weight of each position in $\Omega(i)$ refers to a channel of Y_i. The middle channel always corresponds to the weight of target pixel i itself. Then, we can obtain a weight map $M_i \in \mathbb{R}^{W \times H \times 1}$ where the index positions outside the $\overline{W} \times \overline{H}$ context region are set to a constant. The weight of the normalized attention map $A_i \in \mathbb{R}^{W \times H \times 1}$ at location k is calculated via a softmax function as follows:

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

where $i, j, k \in \{1, 2, ..., W * H\}$ and $M_i = \{m_i^1, m_i^2, ..., m_i^{W*H}\}$. Thus, we obtain a spatial attention map $A_i = \{a_i^1, a_i^2, ..., a_i^{W*H}\}$. The normalized attention weights for all of the locations can be calculated in the same way: $A = \{A_1, A_2, ..., A_{W*H}\}$. Finally, the original features

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2997178, IEEE Transactions on Multimedia

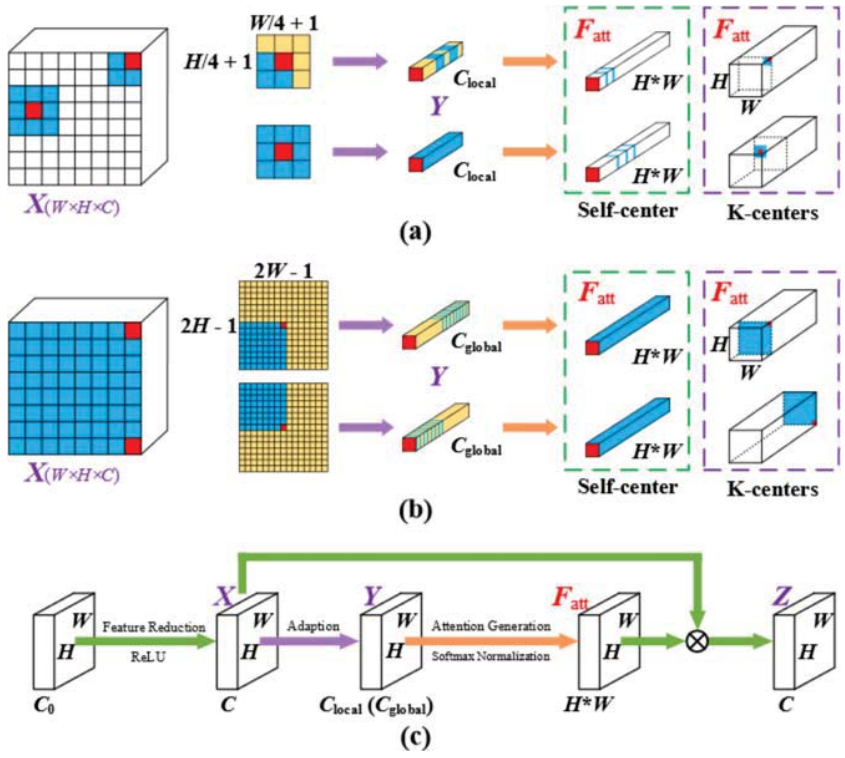

Fig. 4. Detailed structure of the proposed CAM: for the two example pixels, (a) and (b) show the detailed process of generating local and global attention weights F_{att} in the 'self-center' branch and 'k-centers' branch, respectively. (c) shows the common workflow for learning both local and global attention features.

X in channel c are computed by a weighted sum by A_i to generate the new features Z_i:

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

where $c \in \{1, 2, ..., C\}$. Similarly, we can obtain attentive features $Z = \{Z_1, Z_2, ..., Z_{W*H}\}$ that have the same size as X.

In the 'k-centers' branch, our goal is to find a different way to generate the attention weights for all of the pixels. We explore the visual attention based on the knowledge that the information of each position can help the prediction of the other positions [25]. By revisiting the spatial relationship of the target pixel (the red node in X) and its contextual pixels $\forall k \in \Omega(7)$ (the blue nodes in Y), the target pixel can be also treated as a contextual pixel in each $\Omega(k)$. For example, the $7th$ position (the target pixel) is placed in $\Omega(10)$ and $\Omega(12)$, as shown on the right of Fig. 3. Thus, the weight for each contextual pixel k can be also derived from Y_k instead of Y_7, which is the essential difference between the two branches. We follow the row-major order to determine the index number of the $7th$ position in each $\Omega(k)$. Based on this index number, we can obtain a channel of Y_k that indicates the correlation degree of the target pixel and its contextual pixel k. For instance, the weight of the $10th$ position of X is the $3rd$ channel of Y_{10}, and the weight of the $12th$ position of X is the $1st$ channel of Y_{12}. More generally, for each target pixel i, we can we obtain a weight map M^i and then compute the weight of the normalized attention map A^i at location k as follows:

$$a_k^i = \frac{\exp(m_k^i)}{\sum_{j=1}^{W*H} \exp(m_j^i)} \tag{3}$$

where $i, j, k \in \{1, 2, ..., W * H\}$. Finally, the attentive features Z_i in 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 $c \in \{1, 2, ..., C\}$. In the same way, we obtain the attended contextual features $Z = \{Z_1, Z_2, ..., Z_{W*H}\}$ in the 'k-centers' branch.

The 'self-center' and 'k-centers' branches offer two different perspectives to compute the attention weights for target pixel i. Based on the attention weights, target pixel i can not only collect information from its contextual pixels but also propagate information to its contextual pixels when generating attentive features for those contextual pixels. Thus, the CAM can exchange the semantic information between each pixel and its contextual pixels bidirectionally.

2) Attention Generation: In view of the scope of context region Ω, visual attention can be subdivided into local attention and global attention. We empirically adopt a $(W/4 + 1) \times (H/4+1)$ neighborhood as the local context region. The global context region is set to a $(2W-1) \times (2H-1)$ over-completed neighborhood, which guarantees that each pixel can build connections with all of the spatial positions. As Figs. 3 and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2997178, IEEE Transactions on Multimedia

4 show, the context region centered at certain locations may cover some invalid areas (the yellow areas). The channels of Y, which are mapped to those invalid regions, will not participate in the back-propagation learning procedure. To simplify the calculations, we reshape the normalized attention map A_i (or A^i) with size $W \times H$ into the channels of $F_{\text{att}} \in \mathbb{R}^{W \times H \times C_{\text{att}}}$ 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 F_{att} at location i, while the normalized weights in the 'k-centers' branch are mapped to the corresponding spatial locations of F_{att} in channel i. In contrast to the global attention, numerous values of F_{att} in the local attention are constrained to be zero for the purpose of cutting off the information of those positions outside the local neighboring context. To achieve this goal, we fill in these corresponding channels with a large negative value β before the 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 $Z \in \mathbb{R}^{W \times H \times C}$ are obtained in one shot:

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

where $\otimes$ denotes a matrix multiplication.

For the original features 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 means 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) shows the common workflows of the local and global attending operations, which match the respective local and global attention blocks of the 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 attention mechanism, we obtain spatial attention maps for all of the locations of the features X. Finally, the normalized weights F_{att}, which denote the relevance between each location and its contexts, operate on each spatial map of X across all of the channels to generate the new feature representation Z.

In addition, another convolutional layer with 3×3 kernels is used for the feature compression in each branch, as the ablation experiments prove that these dimension-reduced features can contribute to better saliency estimation along with the 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 of 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 connections with the other positions, which will dramatically improve 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 CANet consists of 13 convolutional layers and 5 max-pooling layers from VGG-16. Each 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 the features derived from the 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 use a pyramid-like structure to gradually sharpen the 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 is followed by a 3×3 convolutional layer with 64 channels and a ReLU activation; the result is 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 CANet consists of six modules (see the green blocks in Fig. 2) that generate feature maps with 64 channels, namely, 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. Although this refinement has been adopted in [54] and [55], our attention-guided model generates attentive contextual features in deeper layers for more powerful feature extraction, which is critical to the localization of foreground regions. More specifically, 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 achieve the optimal performance in the current embedding setting, which will be fully investigated in the ablation experiments. On the other hand, if the CAM is continually applied in Dec_3 or beyond that point, the C_{global} in the global attention will be very large, which could sharply increase the computation cost.

In the training process, a one-channel convolutional layer with a 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 Dec_1 to generate the final probability map.

IV. Experiments

A. Dataset and Setup

1) Datasets: Six public saliency datasets, namely, ECSSD [42], HKU-IS [10], DUTS [43], DUT-O [6], PASCAL-S [44] and SOD [45], are selected to evaluate the proposed CANet.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2997178, IEEE Transactions on Multimedia

TABLE I
QUANTITATIVE SOD PERFORMANCE COMPARISONS OVER SIX PUBLIC SOD DATASETS INCLUDING ECSSD [42], HKU-IS [10], DUTS-TE [43], DUT-O [6], PASCAL-S [44], AND SOD [45]. THE BEST MAXF/MAE IS HIGHLIGHTED IN BOLDFACE.

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 [27]	0.907	0.059	0.891	0.052	0.812	0.065	-	-	0.830	0.096	0.823	0.127
DCL [14]	VGG16 [27]	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 [27]	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 [27]	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 [27]	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 [24]	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 [27]	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 [24]	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 [27]	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-R [20]	ResNet50 [24]	0.935	0.046	0.918	0.043	0.860	0.051	0.803	0.065	0.868	0.078	0.853	0.103
PAGENet [49]	VGG16 [27]	0.931	0.042	0.918	0.037	0.838	0.052	0.791	0.062	0.858	0.079	0.837	0.110
AFNet [50]	VGG16 [27]	0.936	0.042	0.923	0.036	0.864	0.046	0.798	0.057	0.870	0.073	0.854	0.108
BASNet [51]	ResNet34 [24]	0.942	**0.037**	0.928	**0.032**	0.860	0.048	0.805	**0.056**	0.860	0.079	0.849	0.112
EGNet [52]	VGG16 [27]	**0.943**	0.041	**0.930**	0.034	**0.877**	**0.044**	0.809	0.057	0.870	0.079	0.859	0.110
Ours	VGG16 [27]	0.938	0.044	**0.930**	0.037	0.876	**0.044**	**0.810**	0.058	**0.880**	0.075	**0.865**	**0.099**

The ECSSD dataset consists of 1,000 natural images, and most of them have semantically meaningful but structurally complex objects. HKU-IS contains 4,447 complex images, and each one 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 are normally utilized for training deep-learning-based models. The corresponding testing dataset (DUTS-TE), which includes 5,019 challenging images, is widely used to compare high-performance models. The DUT-O dataset is a representative SOD benchmark that contains a total of 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, and most of them include cluttered backgrounds or complex salient objects. All of 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 certain binary masks. A binary mask is denoted by $BM \in \{0,1\}$. The precision and recall are obtained by comparing BM and GT as follows: $precision = |BM \cap GT| / |BM|$ and $recall = |BM \cap GT| / |GT|$. For a specific dataset, we compute a series of precision and recall pairs over all of the saliency maps. Then, we employ mean value pairs to draw the PR curve. The second metric is the F-measure score, which evaluates the comprehensive quality of the saliency results. This metric 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 is the MAE (Mean Absolute Error) score, and it is used to measure the dissimilarity between the binary ground truth mask GT and the predicted score map SM in a straight-forward way:

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

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

3) Implementation Details: Our model is implemented using the Caffe [38] library. The training dataset of DUTS [43] is employed to train our saliency network with the standard binary cross entropy loss. Note that all of 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 learning rate that is smaller by a factor of 0.1. Additionally, the learning rate is decreased by a factor of 10 after 10,000 steps. The maximum iteration

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2997178, IEEE Transactions on Multimedia

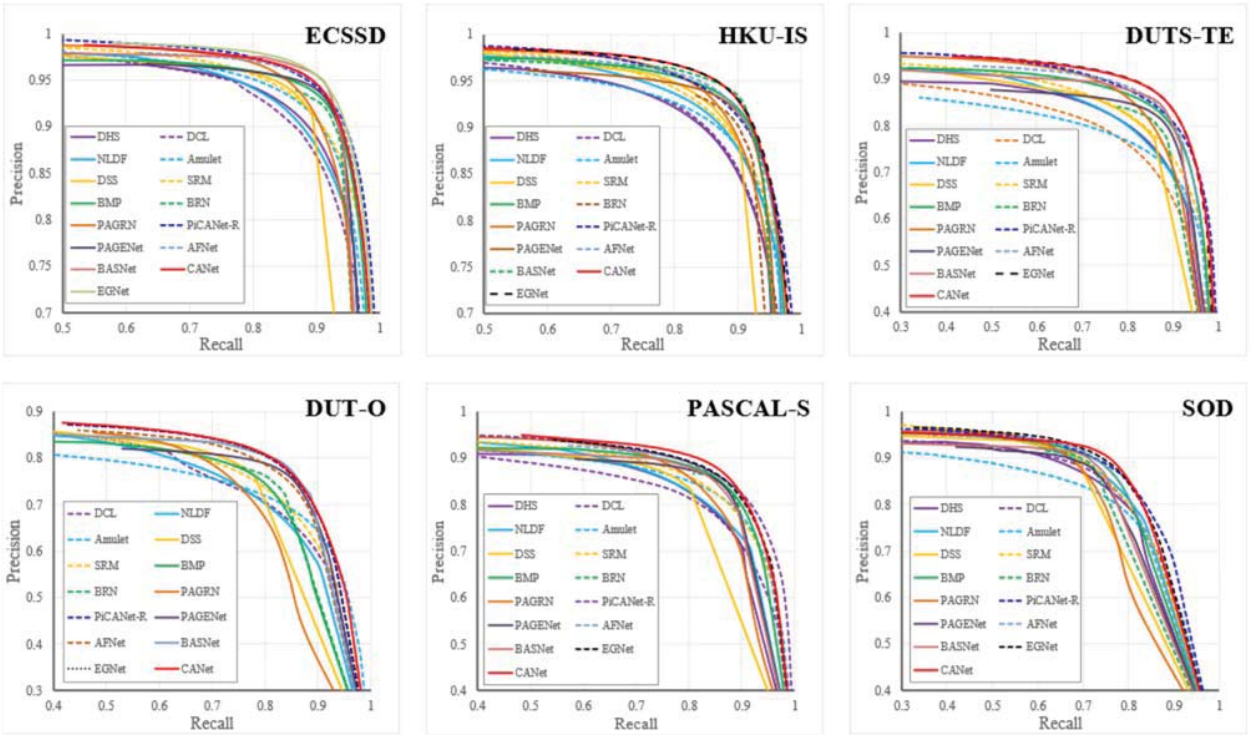

Fig. 5. The precision-recall curves for state-of-the-art methods and our proposed CANet over the six public benchmarking datasets ECSSD, HKU-IS, DUTS-TE, DUT-O, PASCAL-S, and SOD (zoom in for more details).

step is set to be 20,000. It takes approximately 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, as it does not have any extra pre-processing or post-processing steps.

B. Comparison with the State-of-the-Art Models

In this section, we compare CANet with the following 14 state-of-the-art SOD models: DHS [13], DCL [14], NLDF [15], Amulet [16], DSS [17], SRM [23], BMPM [22], DGRL [12], PAGR [19], PiCANet-R [20], PAGENet [49], AFNet [50], BASNet [51], and EGNet [52]. To ensure a fair comparison against these existing methods, we use saliency results that 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 quantitative comparisons with the selected state-of-the-art methods. As Table I and Fig. 5 show, CANet achieves the best F-scores on DUT-O, PASCAL-S, and SOD, which demonstrates its outstanding discriminative capability in detecting salient regions within these challenging datasets. Regarding the MAE scores, CANet achieves the best score over the challenging datasets DUTS-TE and SOD and is on a par with the other methods for the other studied datasets. Fig. 5 shows the precision-recall curves over the six benchmark datasets. It can easily be seen that CANet performs favorably against the other existing models on most of the SOD datasets. Note that CANet achieves comparable performance with respect to the latest SOD model EGNet [52]; e.g., whereas CANet performs slightly better on PASCAL-S and SOD, EGNet performs slightly better on ECSSD.

2) Qualitative Performance Comparison: Fig. 6 shows a qualitative comparison of the proposed CANet with many state-of-the-art SOD models. CANet can consistently highlight salient objects and filter out the redundant background information in various complex scenes, such as images with one or more small objects (rows 1, 2 and 7), large objects (rows 3, 4 and 5), cluttered backgrounds (rows 6 and 9), objects touching the image boundary (rows 3, 4, 5, 6 and 7), low color contrast (rows 1, 4 and 6), and non-center bias (rows 7 and 8). From these visual examples, we make the following two observations. First, some models often suffer from the distractions of complex backgrounds and thereby falsely assign higher saliency values to parts of the non-salient regions. Second, portions of the 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, as it 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 in improving 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 of the experiments are performed on two large-scale

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2997178, IEEE Transactions on Multimedia

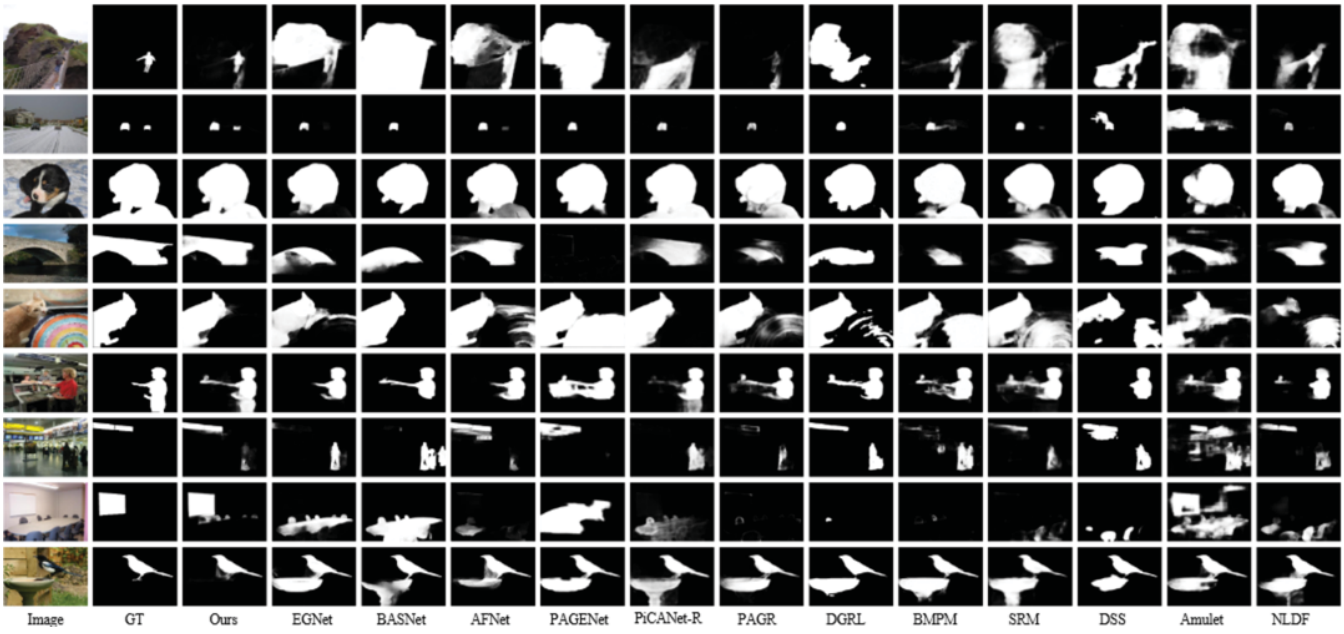

Fig. 6. Qualitative comparisons of our CANet with several state-of-the-art SOD models: CANet consistently produces more accurate predictions of salient objects than the state-of-the-art SOD models EGNet [52], BASNet [51], AFNet [50], PAGENet [49], PiCANet-R [20], PAGR [19], DGRL [12], BMPM [22], SRM [23], DSS [17], Amulet [16], NLDF [15] (zoom in for more details).

TABLE II
ABLATION ANALYSIS OF DIFFERENT DESIGNS IN CANET, INCLUDING THE SELF-CENTER NETWORK BRANCH, K-CENTERS NETWORK BRANCH, FEATURE COMPRESSION, LOCAL ATTENTION AND GLOBAL ATTENTION. THE BEST MAXF AND MAE RESULTS ARE HIGHLIGHTED IN BOLDFACE.

No.	Self-center	k-centers	Feature Compression	Local Attention	Global Attention	DUT-O [6]		DUTS-TE [43]	
						maxF	MAE	maxF	MAE
1	×	×	×	×	×	0.690	0.085	0.738	0.079
2	×	×	√	×	×	0.782	0.063	0.854	0.049
3	√	√	×	√	×	0.797	0.059	0.866	0.047
4	√	√	×	×	√	0.801	0.060	0.866	0.047
5	√	√	√	√	×	0.801	0.058	0.870	0.045
6	√	√	√	×	√	0.806	0.059	0.869	0.045
7	√	√	×	√	√	0.801	0.062	0.868	0.048
8	√	√	√	√	√	**0.810**	**0.058**	**0.876**	**0.044**
9	√	×	√	√	√	0.804	0.058	0.870	0.045
10	×	√	√	√	√	0.805	0.059	0.870	0.046

saliency datasets DUT-O [6] and DUTS-TE [43]. We adopt two common metrics, namely, the maximum F-measure and MAE, to evaluate the effectiveness of the different design options. The quantitative evaluation results are summarized in Table II and Table III. We do not conduct more experiments to show the effectiveness of the deep supervision training strategy, as it has been proven to be effective in many previous works.

1) Local Attention: To validate the effectiveness of local attention, we replace the CAM with other modified versions, as shown in Table II. Note that all of the variants except No. 1 adopt the hierarchical refinement network. The No. 2 setting refers to a common encoder-decoder model without any attention scheme. It can be seen that the hierarchical refinement architecture can improve the boundary quality greatly. The No. 3 setting corresponds to a model that only incorporates local attention. By combining the local dimension-reduced features and the local attentive features, we can obtain a modified module (the No. 5 setting). It is clear that the No. 5 setting consistently achieves better performance than No. 2 and No. 3. This finding indicates that our local attention is useful to extract local information for better saliency inference. Note that these three settings do not consider global attention. To investigate whether local attention is also complementary to global attention, we also add global attention to No. 2 and No. 5 to obtain versions No. 6 and No. 8, respectively. The comparison results further confirm that local attention can boost the overall performance. In terms of the maxF score

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2997178, IEEE Transactions on Multimedia

TABLE III
ABLATION STUDY OF DIFFERENT EMBEDDING CHOICES.

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

TABLE IV
COMPARISON WITH OTHER ATTENTION-BASED METHODS. THE BEST SCORE IS HIGHLIGHTED IN BOLDFACE.

Method	DUT-O [6]		DUTS-TE [43]		Channel-wise Attention	Spatial Attention	FPS
	maxF	MAE	maxF	MAE			
PAGR [19]	0.771	0.071	0.854	0.055	√	√	-
PiCANet-R [20]	0.803	0.065	0.860	0.051	×	√	7
PAGENet [49]	0.791	0.062	0.838	0.052	×	√	25
SE block [53]	0.799	0.059	0.863	0.047	√	×	42
Ours	**0.810**	**0.058**	**0.876**	**0.044**	×	√	32

and MAE score, we observe that No. 6 already outperforms most of the existing models by a large margin, as shown in Table I. It is well known that improving an existing high-performance model is not easy. Nevertheless, our final setting in No. 8 improves upon No. 6 (DUT-O: 0.806 $\rightarrow$ 0.810, and DUTS-TE: 0.869 $\rightarrow$ 0.876) in terms of the maxF score. This improvement demonstrates that the proposed local attention contributes to the saliency estimation.

2) Global Attention: Similar to the various design options in the local attention case, we also study the effect of global attention in two respects: without local attention and with local attention. Table II provides detailed evaluation results of the different configurations. Compared with No. 2 and No. 4, the No. 6 variant has the best performance on both of the large-scale datasets. As can be seen, the combination of global attention and local dimension-reduced features can achieve the highest maxF score on DUT-O when we randomly combine any two components (No. 5, No. 6, and No. 7). The underlying reason is that the largest receptive field from the global attention and the smallest receptive field from the local features may gain a maximum cooperation effect. Furthermore, we also see a large increase in performance when we add global attention to variant No. 5. This result 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 the two branches in the CAM for salient object detection, we retrain another two modified models with the same training dataset and setting. No. 9 represents a model that only has the 'self-center' branch, and No. 10 represents a model that only has the 'k-centers' branch. Our final setting in the CAM corresponds to model No. 8, which contains both parallel branches. The quantitative comparison results of these three variants are depicted in Table II. Clearly, model No. 8 performs favorably against the other two models, which 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. 2 setting in Table II. 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 the input features. More details can be found in Table III. The C_{global} value will be set to be 16,139 if we continue to apply the CAM in Dec_3. However, this setting may greatly increase the computation cost. Thus, to achieve a better tradeoff between the performance and the computation, we only compare three different embedding choices. Even though only Dec_6 is equipped with the CAM, we observe a great improvement upon the baseline model. As shown in Table III, No. 3 is the best embedding choice for our model in terms of the maxF score and MAE score.

5) Comparison with Other Attention-based Models: We perform an extra experiment to further demonstrate the effectiveness of our CAM as compared with the existing attention-based SOD models. Specifically, we replace the CAMs in

CANet with squeeze-and-excitation (SE) modules [53] that focus on the channel-wise attention by filtering out irrelevant feature channels. The experimental results are shown in Table IV. As the table shows, the CAM outperforms the SE modules consistently, as it explicitly generates spatial attention by building connections between each pixel and its neighboring pixels. CANet's efficiency is lower than that of the modified version with SE but higher than those of recent attention-based SOD models.

V. Conclusion

In this paper, we have developed a context-aware attention network for salient object detection. By fully exploring the means of information collection and the scope of the context, our CANet can effectively enhance the ability to identify salient objects and thereby generate accurate prediction results. In detail, with the guidance of the proposed spatial attention mechanism, we can selectively aggregate the contextual information for each pixel and thereby filter out redundant background information and ensure a high degree of completeness. By learning the 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, 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 the saliency results. Exhaustive experiments with 14 state-of-the-art methods on six saliency datasets demonstrate the superior performance of our CANet. In the future, we intend to improve the 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] S. Lu and J.H. Lim, "Saliency Modeling from Image Histograms," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 321-332.

[36] S. Lu, C. Tan, and J.-H. Lim, "Robust and efficient saliency modeling from image co-occurrence histograms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 195–201, Jan. 2014.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2997178, IEEE Transactions on Multimedia

[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] 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.
[48] T. Zhao, and X. Wu, "Pyramid Feature Attention Network for Saliency Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3085-3094.
[49] W. Wang, S. Zhao, J. Shen, S. Hoi, and A. Borji, "Salient Object Detection With Pyramid Attention and Salient Edges," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1448-1457.
[50] M. Feng, H. Lu, and E. Ding, "Attentive Feedback Network for Boundary-Aware Salient Object Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1623-1632.
[51] X. Qin et al., "BASNet: Boundary-Aware Salient Object Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7479-7489.
[52] J. Zhao et al., "EGNet: Edge Guidance Network for Salient Object Detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8779-8788.
[53] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7132-7141.
[54] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1395–1403.
[55] Y. Liu, M.-M. Cheng, X. Hu, K. Wang, and X. Bai, "Richer convolutional features for edge detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3000-3009.
[56] 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.
[57] 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.

Qinghua Ren received the M.S. degree in electrical engineering from Southeast University, Nanjing, China, in 2016. He is currently pursuing the Ph.D. degree under the supervision of Prof. Renjie Hu in Southeast University. He is currently a visiting student in the School of Computer Science and Engineering at Nanyang Technological University, Singapore. His research interests include computer vision, object detection and deep learning.

Shijian Lu received his PhD in electrical and computer engineering from the National University of Singapore. He is currently an Assistant Professor with School of Computer Science and Engineering, the Nanyang Technological University, Singapore. His major research interests include image and video analytics, visual intelligence, and machine learning. He has published more than 100 international refereed journal and conference papers and co-authored over 10 patents in these research areas. He is currently an Associate Editor for the journal Pattern Recognition (PR). He has also served in the program committee of a number of conferences, e.g. the Area Chair of the International Conference on Document Analysis and Recognition (ICDAR) 2017 and 2019, the Senior Program Committee of the International Joint Conferences on Artificial Intelligence (IJCAI) 2018 and 2019, etc.

Jinxia Zhang received the B.S. degree in the Department of Computer Science and Engineering, Nanjing University of Science and Technology, China in 2009 and the PhD degree in the Department of Computer Science and Engineering, Nanjing University of Science and Technology, China in 2015. She was a visiting scholar in Visual Attention Lab at Brigham and Women's Hospital and Harvard Medical School from 2012 to 2014. She is currently an associate professor in the School of Automation, Southeast University. Her research interests include saliency detection, knowledge transfer, computer vision and deep learning.

Renjie Hu received the Ph.D. degree in electrical engineering from Southeast University, Nanjing, China, in 2002. He is currently a Professor at Southeast University and serves as the Chief of Electrical and Electronic Experiment Center. His research interests include power electronics, power quality management, and image processing.