

SAC-Net: Spatial Attenuation Context for Salient Object Detection

Xiaowei Hu, Chi-Wing Fu, Lei Zhu, Tianyu Wang, and Pheng-Ann Heng

arXiv:1903.10152v3 [cs.CV] 12 May 2020

Abstract—This paper presents a new deep neural network design for salient object detection by maximizing the integration of local and global image context within, around, and beyond the salient objects. Our key idea is to adaptively propagate and aggregate the image context features with variable attenuation over the entire feature maps. To achieve this, we design the spatial attenuation context (SAC) module to recurrently translate and aggregate the context features independently with different attenuation factors and then to attentively learn the weights to adaptively integrate the aggregated context features. By further embedding the module to process individual layers in a deep network, namely SAC-Net, we can train the network end-to-end and optimize the context features for detecting salient objects. Compared with 29 state-of-the-art methods, experimental results show that our method performs favorably over all the others on six common benchmark data, both quantitatively and visually.

Index Terms—Spatial attenuation context, salient object detection, saliency detection, deep learning.

I. INTRODUCTION

Salient object detection aims to distinguish the most visually distinctive objects from an input image and it is an effective pre-processing step in many image processing and computer vision tasks, *e.g.*, object segmentation [1] and tracking [2], video compression [3] and abstraction [4], image editing [5], texture smoothing [6], as well as few-shot learning [7]. It is a fundamental problem in computer vision research and has been extensively studied in the past decade.

Early works attempt to detect salient objects based on low-level cues like contrast, color, and texture [8], [9], [10], [11]. However, relying on low-level cues is clearly inadequate to finding salient objects, which involve high-level semantics. Hence, most recent methods [12], [13], [14], [15], [16] employ convolutional neural networks (CNNs) and take a data-driven approach to the problem by leveraging both high-level semantics and low-level details extracted from multiple CNN layers [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33]. However, since the convolution operator in CNN processes a local neighborhood in the spatial domain [34], existing methods tend to miss global spatial semantics in the results, *e.g.*, they may misrecognize background noise as salient objects; see Section IV-B for quantitative and qualitative comparisons.

X. Hu, L. Zhu, and T. Wang are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China.

C.-W. Fu and P.-A. Heng are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China and also with Shenzhen Key Laboratory of Virtual Reality and Human Interaction Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China.

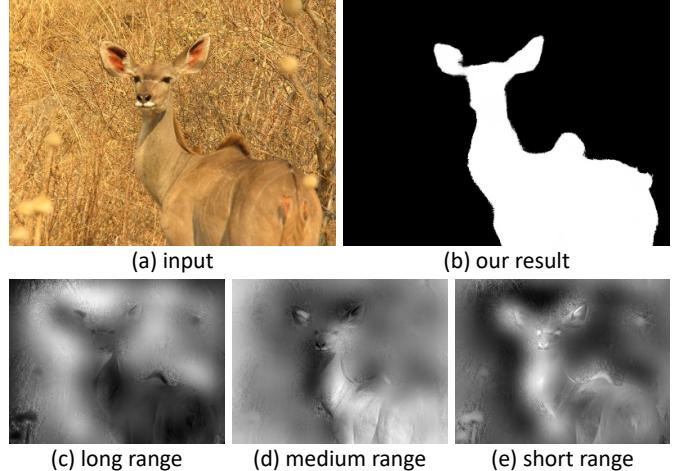


Fig. 1: A challenging example (a), where our method is still able to find the object contour (b); see (c)-(e) for the attention weights learned for the context with different spatial ranges.

Essentially, salient objects are key elements that stand out from the background. Such an inference process [35] should involve not only the local image context within and around the salient objects, but also the global image context, as well as a suitable integration of the various context features. Ideally, after extracting context features per image pixel, if we can connect all these features and let them communicate with every other over the spatial domain, we can optimize the feature integration for maximized performance. However, it is computationally infeasible in practice. Hence, we present to propagate context features with different attenuation factors over the spatial domain of the image and learn to aggregate the resulting features adaptively; by then, our network can learn to detect salient objects by adaptively considering context features within, around, and even far from, the salient objects.

To achieve this, we present the spatial attenuation context, which is achieved by the following steps: (i) the image context is aggregated by propagating the information pixel by pixel over the whole feature maps; as a result, each pixel will obtain the global information from all other pixels of the feature maps; (ii) the propagation ability is affected by an attenuation factor, where a large attenuation factor reduces the information propagation and leads to a short-range context while a small attenuation factor improves the information propagation and leads to a long-range context; (iii) the image contexts with different ranges are dynamically merged by learning a set of attention weights. Fig. 1 shows a challenging example

with the associated attention maps learned in our network for *integrating the various image context*: (c) long-range context aggregated with a small attenuation factor helps locate the global background; (d) medium-range context helps identify the image regions of the same object; and (e) short-range context aggregated with a large attenuation factor helps locate the boundary between salient and non-salient regions. Please see the supplementary material for detailed explanations with more illustrations.

In details, we formulate the *spatial attenuation context module*, or *SAC module* for short, in a deep network to allow the image features in a CNN to *propagate over variable spatial ranges by articulating different attenuation factors in the propagation*. Our module has two rounds of recurrent translations to propagate and aggregate the image features. In each round, we propagate features independently using different attenuation factors towards different directions in the spatial domain; further, we formulate an attention mechanism to learn the weights to combine the aggregated features. Hence, we can adopt different attenuation factors (or influence ranges) for different image features. Furthermore, we deploy an SAC module in each layer of our network and predict a saliency map per layer based on the output from the SAC module and the convolutional features. Below, we summarize the major contributions of this work:

- We design the spatial attenuation context (SAC) module to recurrently propagate the image features over the whole feature maps with variable attenuation factors and learn to adaptively integrate the features through an attention mechanism in the module. Then, we adopt the SAC module in each layer of our network architecture to learn the spatial attenuation context in different layers, and train the whole network in an end-to-end manner for salient object detection.
- We evaluate our method and compare it against 29 state-of-the-art methods on six common benchmark data. Results show that our method performs favorably over all the others for all the benchmark data. Our code, trained models, and predicted saliency maps are publicly available at <https://xw-hu.github.io/>.

II. RELATED WORK

Rather than being comprehensive, we discuss mainly the methods on single-image salient object detection. Early methods use hand-crafted priors such as image contrast [9], [36], color [37], [38], texture [39], [40], and other low-level visual cues [41]; see [42] for a survey. Clearly, hand-crafted features are insufficient to capture high-level semantics, so methods based on them often fail for nontrivial inputs.

Recent works [12], [13], [16] exploit convolutional neural networks (CNN) to learn deep features for detecting salient objects. However, since these methods just take features at deep CNN layers, they tend to miss the details in the salient objects, which are captured mainly in the shallow layers. Several recent works [14], [15], [21], [19], [24], [20], [18], [23], [25], [22], [17], [43], [26] enhance the detection quality by further integrating features in multiple CNN layers to

simultaneously leverage more global and local context in the inference process. Among them, Li *et al.* [21] explored the semantic properties and visual contrast of salient objects, Hou *et al.* [19] created short connections to integrate features in different layers, while Zhang *et al.* [24] derived a resolution-based feature combination module and a boundary-preserving refinement strategy. Hu *et al.* [20] recurrently aggregated deep features to exploit the complementary saliency information between the multi-level features and the features at each individual layer. Later, Deng *et al.* [18] adopted residual learning to alternatively refine features at deep and shallow layers. Zhang *et al.* [23] formulated a bi-directional message passing model to select features for integration. Zhang *et al.* [25] designed an attention-guided network to progressively select and integrate multi-level information. Li *et al.* [22] used a two-branch network to simultaneously predict the contours and saliency maps. Chen *et al.* [17] leveraged residual learning and reverse attention to refine the saliency maps. Li *et al.* [33] presented a contrast-oriented deep neural network, which adopts two network streams for both dense and sparse saliency inference. Zhang *et al.* [26] designed a symmetrical CNN to learn the complementary saliency information and presented a weighted structural loss to enhance the boundaries of salient objects. Wang *et al.* [43] explored the global and local spatial relations in deep networks to locate salient objects and refine the object boundary. Although the detection quality keeps improving, the exploration of global spatial context, particularly in the shallow layers, is still heavily limited by the convolution operator in CNN, which is essentially a local spatial filter [34].

Very recently, Liu and Han [44] incorporated global context and scene context by developing a deep spatial long short-term memory model. Liu *et al.* [45] aggregated the attended contextual features from a global/local view in feature maps of varying resolutions. Wang *et al.* [27] presented a pyramid attention structure and leveraged the salient edge information to better segment salient objects. Feng *et al.* [46] designed an attentive feedback network to further explore the boundaries of the salient objects. Zhao and Wu [28] used the dilated convolution and channel-wise & spatial attention to aggregate multi-scale context features. Wu *et al.* [29] proposed to discard the feature maps at shallow layers for acceleration and used the saliency map generated from one network branch to refine the features of another branch. Liu *et al.* [30] introduced two pooling-based modules to progressively refine the highly semantic features for detail enriched saliency maps. Wang *et al.* [31] predicted the saliency maps by iteratively aggregating the feature maps in the top-down and bottom-up manner. Zhang *et al.* [47] incorporated the semantic information of salient objects from the image captions. Qin *et al.* [48] formulated a boundary-aware salient object detection network by combining a deeply supervised encoder-decoder and a residual refinement module, and leveraged a hybrid loss to optimize the whole network. Wu *et al.* [32] jointly performed foreground contour detection and edge detection tasks by using multi-task intertwined supervision. Fu *et al.* [49] presented a Deepside to incorporate hierarchical CNN features and fused multiple side outputs based on a segmentation-based pooling. Li *et al.* [50]

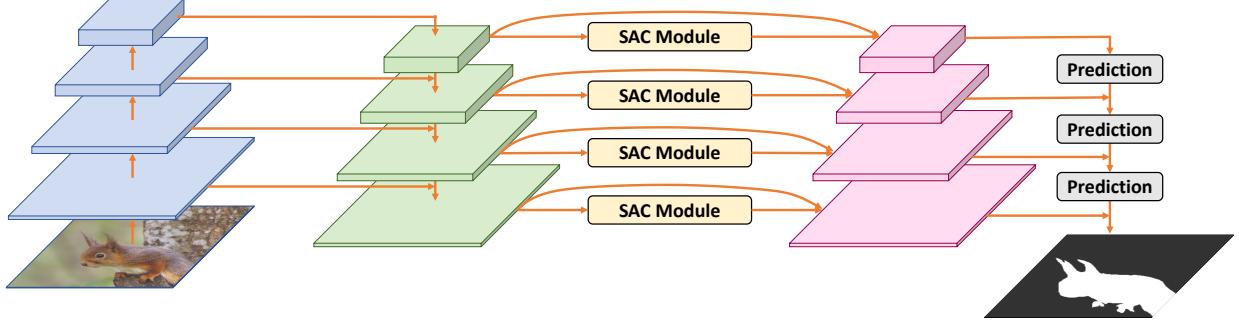


Fig. 2: The schematic illustration of our spatial attenuation context network (SAC-Net): (i) extract feature maps (in blue) in different resolutions from the input image using a convolutional neural network; (ii) construct a feature pyramid (in green) by successively upsampling the feature map at a deep layer and combining the upsampled result with the feature map at an adjacent shallower layer; (iii) use SAC modules (see Fig. 4) to generate spatial attenuation context features for each layer; (iv) concatenate the outputs from the SAC modules with the convolutional features (in red); and (v) lastly, successively predict a saliency map at each layer and take the final saliency map of the largest resolution as the network output. In the figure, feature maps are indicated by blocks and thicker blocks of smaller sizes are higher-level features at deeper layers.

developed a multiscale saliency refinement network, which is used for instance-level salient object segmentation. Zhu *et al.* [51] learned the attentional dilated features to detect the salient objects. Even the detection performance continues to improve on the benchmarks [52], [53], [54], [55], [39], [40], current methods may still miss local parts in salient objects and misrecognize noises in non-salient regions as salient objects. Except the above works, Song *et al.* [56] presented a novel multi-scale attention network for accurate object detection. Peng *et al.* [57] proposed two-stream collaborative learning with a spatial-temporal attention approach for video classification. He *et al.* [58] developed a multi-scale and multi-granularity deep reinforcement learning approach for fine-grained visual categorization. Peng *et al.* [59] formulated an object-part attention model for weakly supervised fine-grained image classification.

The recent works [44], [45] that emphasize the importance of reasoning spatial context for salient object detection. Comparing with the PiCANet [44], [45], which aggregates the global context formation on the feature maps with small resolutions by adopting the expensive long short-term memory models and aggregates the local context information on the feature maps with large resolutions through convolutions, we leverage and selectively aggregate surrounding image context spatially in the same CNN layer by a new concept, i.e., spatial attenuation context, which attentively allows the context features to recurrently translate with varying attenuation factors (including local and global information) on the feature maps with any resolutions.

III. METHODOLOGY

Fig. 2 outlines the architecture of our spatial attenuation context network (SAC-Net), which takes a whole image as input and predicts the saliency map in an end-to-end manner. First, we use a CNN to generate feature maps in different resolutions and progressively propagate the image features at deep layers to feature maps at shallow layers to construct a feature pyramid [60]. After that, we use our SAC modules to

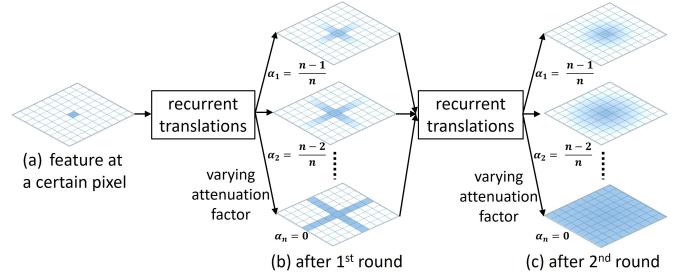


Fig. 3: Illustrating how the image features propagate with varying attenuation factors (α_k) inside the SAC module; please see Fig. 4 for the detailed module architecture.

harvest spatial attenuation context per layer and concatenate the module outputs with the corresponding convolutional features. Lastly, we predict a result per layer, upsample and merge it with the shallower-layer output, and take the result of the largest resolution as the final network output. In the following subsections, we first elaborate on the SAC module, and then present the strategies to train and test our network for salient object detection.

A. Spatial Attenuation Context Module

Fig. 4 shows the architecture of the *spatial attenuation context module*, or *SAC module*, which takes a feature map as input and produces spatial attenuation context in the same resolution. As presented earlier, the spatial attenuation context contains image context aggregated by propagating local image context using varying attenuation factors via an attention mechanism; hence, we can disperse the local image context adaptively over the whole feature maps.

See again the SAC module in Fig. 4. First, we use a 1×1 convolution on the input feature map to reduce the number of feature channels. Then, we adopt recurrent translations with varying attenuation factors (α_k) to disperse the local image features in four different directions; see the illustration in Fig. 3(b) & the detailed structure of recurrent translations in

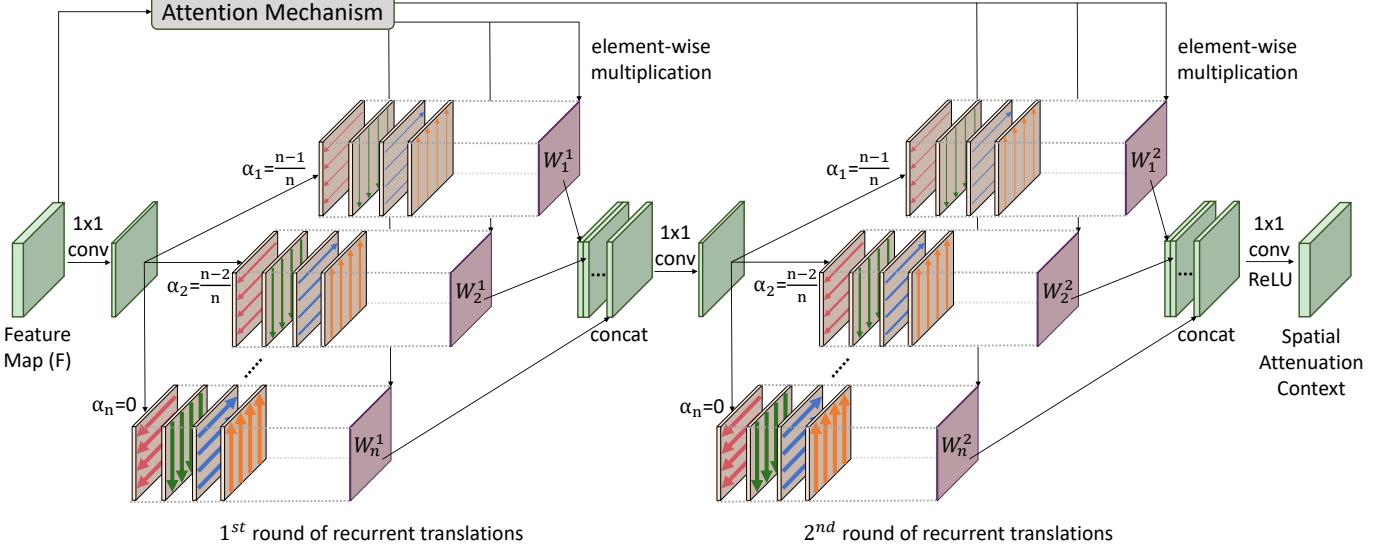


Fig. 4: The schematic illustration of the spatial attenuation context (SAC) module. We adopt two rounds of recurrent translations to propagate and aggregate image features. In each round, the colored arrows show the recurrent translation direction, while thicker (or thinner) arrows indicate stronger (or weaker) information propagation with less (or more) attenuation.

Fig. 4. At this moment, each pixel learns the spatial attenuation context along the four directions. After two rounds of recurrent translations, we adaptively disperse the local features over the 2D domain; see Fig. 3(c). Hence, each pixel knows *the global spatial attenuation context over the entire feature map*. More importantly, we learn the weights to combine the recurrently-aggregated results via an attention mechanism in an end-to-end manner (Fig. 4), so each pixel in the SAC module output can receive spatial context *adaptively* from its surroundings; please see the supplementary material for the detail explanations.

Recurrent-attenuating translation. To optimize the dispersal of local context, we first formulate a parametric model to recurrently aggregate the image features with attenuation. Given the feature map after a 1×1 convolution (see Fig. 4), we recurrently translate its features using different attenuation factors α_k in four principal directions: left, up, right, and down. Moreover, to ensure manageable memory consumption, we set the number of feature channels in each recurrently-aggregated feature map as $\lfloor \frac{256}{n} \rfloor$, where n is the number of different attenuation factors in the SAC module; see Table V for an experiment on n .

Denoting $f_{i,j}$ as the feature at pixel (i, j) in a feature map, our recurrently-attenuating translation process propagates features progressively over the spatial domain using the following equation (typically in the up direction) :

$$\begin{aligned} f_{i,j}^{up}(\alpha_k, \beta) &= \max(r_{i,j}^{up}, 0) + \beta \min(r_{i,j}^{up}, 0) \\ \text{and } r_{i,j}^{up} &= (1 - \alpha_k) \cdot f_{i-1,j}^{up} + f_{i,j}, \end{aligned} \quad (1)$$

where $\alpha_k = \frac{n-k}{n}$ ($k \in \{1, 2, \dots, n\}$) is the attenuation factor, and β is a learnable parameter in our recurrently-attenuating translation model.

In Eq. (1), we recurrently aggregate image features by using $r_{i,j}^{up}$, where a smaller α_k (close to zero) allows the features to propagate over a longer distance, while a larger α_k (close

to one) limits the propagation, so the related local features affect a smaller local area; see again the illustration in Fig. 3. Moreover, when $r_{i,j}^{up} < 0$, the first term in $f_{i,j}^{up}$ will become zero, and β will be multiplied with $r_{i,j}^{up}$. We define β in Eq. (1) to reduce the feature magnitude when it is negative. Since we learn the value of β for each feature channel, we can introduce nonlinearities when aggregating the spatial context and express more complex relations among the local features. Note that in our experiments, we initialize β as 0.1 for all the feature channels and learn it automatically during the network training process; in practice, we found that β rarely goes beyond one in our experiments.

Attention mechanism. After recurrently-translating the input feature map using different attenuation factors in four directions, we will obtain $4n$ feature maps; see the feature maps with colored arrows in Fig. 4. As discussed earlier, the long-range image context reveals global semantics, while the short-range context helps identify the boundary between salient and non-salient regions. To adaptively leverage the complementary advantages of these aggregated spatial context features, we formulate an attention mechanism to learn the weights for selectively integrating them.

As shown at the top left corner in Fig. 4, we take the input feature map F as the input to the attention mechanism and produce a set of unnormalized attention weights $\{A_1^1, A_2^1, \dots, A_n^1\}$, each corresponding to a particular attenuation factor; superscript 1 indicates that these weights are for the first round of recurrent translations. Then, we apply the Softmax function (Eq. (3)) to normalize the weights and produce the attention weight maps $\{W_1^1, W_2^1, \dots, W_n^1\}$ associated with different attenuation factors (see Fig. 4):

$$\{A_1^1, A_2^1, \dots, A_n^1\} = \Psi(F; \theta), \text{ and} \quad (2)$$

$$w_{i,j,k}^1 = \frac{\exp(a_{i,j,k}^1)}{\sum_k \exp(a_{i,j,k}^1)}, \quad (3)$$

TABLE I: Comparing our method (SAC-Net) with 29 state-of-the-art methods using the F_β , S_m and MAE metrics. Top two results are highlighted in **red** and **blue**, respectively; “-” indicates results that are not publicly available on the corresponding dataset; and “**” indicates CRF is used as a post-processing step in the methods.

Dataset	-	ECSSD [39]			PASCAL-S [53]			SOD [61]			HKU-IS [52]			DUT-OMRON [40]			DUTS-test [55]		
		F_β	S_m	MAE	F_β	S_m	MAE	F_β	S_m	MAE	F_β	S_m	MAE	F_β	S_m	MAE	F_β	S_m	MAE
SAC-Net* (ours)	-	0.954	0.930	0.028	0.876	0.801	0.070	0.884	0.801	0.092	0.945	0.925	0.023	0.832	0.846	0.050	0.898	0.878	0.032
PiCA-RC* [45]	2018	0.940	0.916	0.035	0.870	0.789	0.073	0.867	0.780	0.094	0.929	0.905	0.031	0.828	0.826	0.054	0.871	0.849	0.040
R ³ Net* [18]	2018	0.935	0.910	0.040	0.845	0.749	0.100	0.847	0.761	0.124	0.916	0.900	0.036	0.805	0.817	0.063	0.833	0.823	0.058
GNLB* [6]	2018	0.931	0.900	0.045	0.840	0.758	0.096	0.837	0.744	0.127	0.917	0.886	0.037	0.800	0.817	0.058	0.830	0.811	0.058
RADF* [20]	2018	0.924	0.894	0.049	0.832	0.754	0.102	0.835	0.759	0.125	0.914	0.889	0.039	0.789	0.815	0.060	0.819	0.814	0.061
DSS* [19]	2017	0.916	0.882	0.053	0.829	0.739	0.102	0.842	0.746	0.118	0.911	0.881	0.040	0.771	0.790	0.066	0.825	0.812	0.057
DCL* [21]	2016	0.898	0.868	0.071	0.822	0.783	0.108	0.832	0.745	0.126	0.904	0.861	0.049	0.757	0.771	0.080	0.782	0.795	0.088
SAC-Net (ours)	-	0.951	0.931	0.031	0.879	0.806	0.070	0.882	0.809	0.093	0.942	0.925	0.026	0.830	0.849	0.052	0.895	0.883	0.034
PoolNet-R [30]	2019	0.944	0.921	0.039	0.865	0.794	0.080	0.869	0.801	0.100	0.934	0.912	0.033	0.830	0.836	0.056	0.886	0.871	0.040
BASNet [48]	2019	0.942	0.916	0.037	0.858	0.785	0.084	0.851	0.772	0.112	0.929	0.909	0.032	0.811	0.836	0.056	0.860	0.853	0.047
CPD-R [29]	2019	0.939	0.918	0.037	0.861	0.789	0.078	0.859	0.771	0.110	0.925	0.906	0.034	0.797	0.825	0.056	0.865	0.858	0.043
AFNet [46]	2019	0.935	0.917	0.042	0.866	0.792	0.076	-	-	-	0.925	0.905	0.036	0.820	0.826	0.057	0.867	0.855	0.045
MLMSNet [32]	2019	0.930	0.909	0.045	0.858	0.790	0.079	0.862	0.790	0.106	0.922	0.906	0.039	0.793	0.809	0.064	0.854	0.851	0.048
CapSal [47]	2019	-	-	-	0.868	0.769	0.079	-	-	-	0.889	0.849	0.057	-	-	-	0.845	0.808	0.060
PiCA-R [45]	2018	0.935	0.917	0.047	0.868	0.800	0.078	0.864	0.793	0.103	0.919	0.904	0.043	0.820	0.832	0.065	0.863	0.859	0.050
ASNet [62]	2018	0.932	0.915	0.047	0.869	0.794	0.075	0.859	0.800	0.105	0.922	0.906	0.041	-	-	-	0.835	0.834	0.060
R ³ Net [18]	2018	0.929	0.910	0.051	0.842	0.761	0.103	0.839	0.770	0.131	0.914	0.897	0.046	0.802	0.819	0.073	0.831	0.829	0.067
BDMPM [23]	2018	0.928	-	0.044	0.862	-	0.074	0.851	-	0.106	0.920	-	0.038	-	-	0.850	-	0.049	
PAGRN [25]	2018	0.927	0.889	0.061	0.849	0.749	0.094	-	-	-	0.918	0.887	0.048	0.771	0.775	0.071	0.854	0.825	0.055
GNLB [6]	2018	0.926	0.904	0.056	0.841	0.772	0.099	0.834	0.762	0.133	0.909	0.891	0.048	0.800	0.824	0.067	0.821	0.822	0.068
DGRL [43]	2018	0.925	0.906	0.045	0.850	0.796	0.080	0.846	0.777	0.104	0.914	0.897	0.037	0.779	0.810	0.063	0.834	0.836	0.051
RAS [17]	2018	0.916	0.893	0.058	0.842	0.735	0.122	0.847	0.767	0.123	0.913	0.887	0.045	0.785	0.814	0.063	0.831	0.828	0.059
C2S [22]	2018	0.911	0.896	0.053	0.845	0.793	0.084	0.821	0.763	0.122	0.898	0.889	0.046	0.759	0.799	0.072	0.811	0.822	0.062
SRM [63]	2017	0.917	0.895	0.054	0.847	0.782	0.085	0.839	0.746	0.126	0.906	0.888	0.046	0.769	0.798	0.069	0.827	0.825	0.059
Amulet [24]	2017	0.913	0.894	0.059	0.828	0.794	0.095	0.801	0.755	0.146	0.887	0.886	0.053	0.737	0.781	0.083	0.778	0.796	0.085
UCF [15]	2017	0.910	0.883	0.078	0.821	0.792	0.120	0.800	0.763	0.164	0.886	0.875	0.073	0.735	0.758	0.131	0.771	0.777	0.117
NLDF [13]	2017	0.905	0.875	0.063	0.831	0.756	0.099	0.810	0.759	0.143	0.902	0.879	0.048	0.753	0.770	0.080	0.812	0.815	0.066
DHSNet [64]	2016	0.907	0.884	0.059	0.827	0.752	0.096	0.823	0.752	0.127	0.892	0.870	0.052	-	-	-	0.807	0.811	0.067
RFCN [14]	2016	0.898	0.860	0.097	0.827	0.793	0.118	0.805	0.717	0.161	0.895	0.859	0.079	0.747	0.774	0.095	0.784	0.791	0.091
ELD [65]	2016	0.867	0.841	0.080	0.771	-	0.121	0.760	-	0.154	0.844	-	0.071	0.719	0.751	0.091	0.738	0.719	0.093
MDF [52]	2015	0.831	0.764	0.108	0.759	0.692	0.142	0.785	0.674	0.155	-	-	-	0.694	0.703	0.092	0.730	0.723	0.094
LEGS [66]	2015	0.827	0.787	0.118	0.756	0.682	0.157	0.707	0.661	0.215	0.770	-	0.118	0.669	-	0.133	0.655	-	0.138
BSCA [67]	2015	0.758	0.725	0.183	0.666	0.633	0.224	0.634	0.622	0.266	0.723	0.700	0.174	0.616	0.652	0.191	0.597	0.630	0.197
DRFI [9]	2013	0.786	-	0.164	0.698	-	0.207	0.697	-	0.223	0.777	-	0.145	-	-	0.647	-	0.175	
SAC-Net (Res50)	-	0.945	0.924	0.034	0.871	0.805	0.072	0.872	0.804	0.093	0.936	0.920	0.028	0.808	0.832	0.057	0.881	0.873	0.037

where $a_{i,j,k}^1 \in A_k^1$ is the unnormalized attention weight at pixel (i, j) for attenuation factor α_k , $w_{i,j,k}^1 \in W_k^1$ are the normalized attention weights, and θ denotes the parameters learned by Ψ , which consists of two 3×3 convolution layers and one 1×1 convolution layer, and we apply the group normalization [68] and ReLU non-linear operation [69] after the first two convolution layers.

Next, we multiply W_k^1 with the corresponding context features aggregated after the recurrent translations:

$$f_{i,j} = \bigoplus_{k=1}^n \left[\begin{array}{l} (f_{i,j}^{up}(\alpha_k, \beta) \oplus f_{i,j}^{down}(\alpha_k, \beta) \\ \oplus f_{i,j}^{left}(\alpha_k, \beta) \oplus f_{i,j}^{right}(\alpha_k, \beta)) \times w_{i,j,k}^1 \end{array} \right], \quad (4)$$

where \times denotes an element-wise multiplication, \oplus denotes the concatenation operator, and $\bigoplus_{k=1}^n$ concatenates all the feature maps for different attenuation factors, after the feature maps are multiplied with the attention weights $(w_{i,j,k}^1)$ by broadcasting the W_k^1 in a channel-wise manner. With the attention weights learned to select and integrate the context features aggregated with different attenuation factors (see again Fig. 1), our network can adaptively control the feature integration and allow the context features to be implicitly dispersed over varying spatial ranges.

Completing the SAC module. After concatenating the features, we complete the first round of recurrent translations

in our SAC module and further apply a 1×1 convolution to reduce the feature channels. Then, we repeat the same process in the second round of recurrent translation using another set of attention weights $\{W_1^2, W_2^2, \dots, W_n^2\}$, which are also learnt through the attention mechanism; see again Fig. 4. After two rounds of recurrent translations, each pixel can obtain context features from the global domain adaptively aggregated with different attenuations; see Fig. 3(c). In the end, we further perform a 1×1 convolution followed by the group normalization and ReLU non-linear operation on the integrated features to produce the SAC module output, i.e., the spatial attenuation context. Since two rounds of recurrent translations are able to obtain the global information, we set the number of rounds as two during the experiments.

B. Training and Testing Strategies

We built our SAC-Net on ResNet-101 [70] and used the feature pyramid network (FPN) [60] (green blocks in Fig. 2) to enhance the feature’s expressiveness. Like [60], we set the channel number of each FPN or SAC layer as 256 and did not use the feature maps at the first layer in both the FPN or SAC module due to the large memory footprint. This framework was implemented based on Caffe [71].

Objective function. We used the cross-entropy loss to train the network. Since we have multiple predictions over different

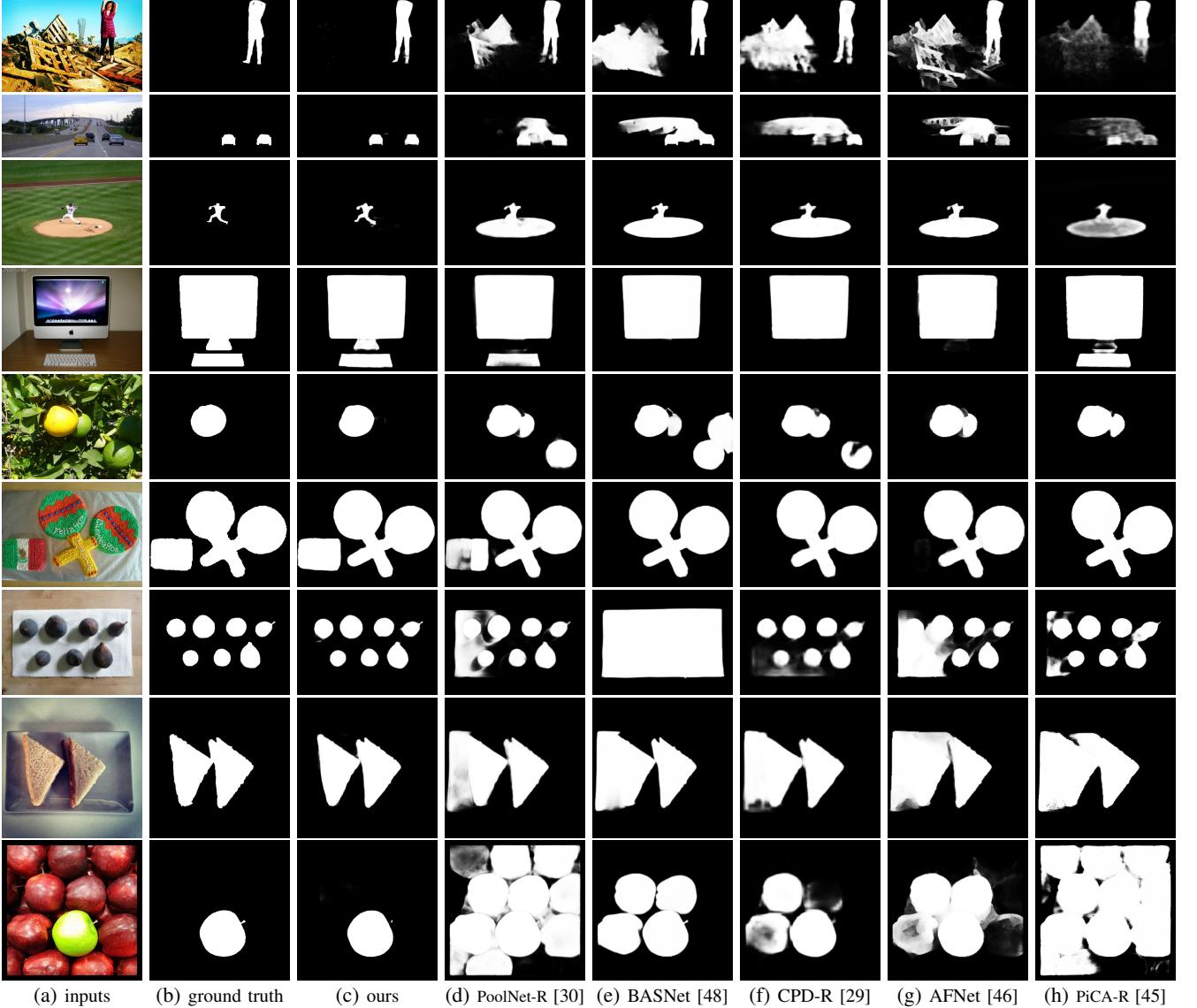


Fig. 5: Visual comparison of saliency maps (c)-(h) produced by different methods. Apparently, our method produces more accurate saliency maps. Results are shown before using CRF.

layers (from deep to shallow) in our SAC-Net (see Fig. 2), the total loss L is defined as the summation of the cross-entropy loss over all the predicted saliency maps:

$$L = - \sum_l \sum_{i,j} g_{i,j} \log(p_{i,j}^l) - (1 - g_{i,j}) \log(1 - p_{i,j}^l), \quad (5)$$

where l is the layer index in network, $g_{i,j}$ is the ground truth value at pixel (i,j) (i.e., one for salient regions, and zero, otherwise), and $p_{i,j}^l \in [0, 1]$ is the predicted saliency value at pixel (i,j) on the result in the network's l -th layer.

Training parameters. We initialized the feature extraction part in our network (frontal blue blocks in Fig. 2) using weights of ResNet-101 [70] trained on ImageNet [72], and initialized other network parts using random noise. Moreover, we adopted two different training strategies to optimize the network. First, we used stochastic gradient descent (SGD) with a momentum value of 0.9 and a weight decay of 0.0005, and we set the learning rate as 10^{-8} , adjusted it to be 10^{-9}

after 13,000 training iterations, and stopped the training after 20,000 iterations. Second, following [30], we used Adam [73] with the first momentum value of 0.9, second momentum value of 0.999, and weight decay of 5×10^{-4} . We set the learning rate as 10^{-5} and stopped the training after 50,000 iterations. The first training strategy is fast while the second strategy achieves better results; see Section IV-C. Also, we horizontally flipped the input images for data argumentation in both training strategies. Lastly, we trained the network on a single NVidia Titan Xp GPU with a mini-batch size of one and updated the weights in every ten training iterations.

Inference. We took the highest-resolution prediction as the overall result and refined the salient object boundary using fully-connected conditional random field (CRF) [74].

TABLE II: Comparing our method (SAC-Net) with the state-of-the-art methods using ResNet-101 as the backbone network. Results are reported before using CRF.

Dataset	-	ECSSD [39]			PASCAL-S [53]			SOD [61]			HKU-IS [52]			DUT-OMRON [40]			DUTS-test [55]		
Metric	Year	F_β	S_m	MAE	F_β	S_m	MAE	F_β	S_m	MAE	F_β	S_m	MAE	F_β	S_m	MAE	F_β	S_m	MAE
SAC-Net (ours)	-	0.951	0.931	0.031	0.879	0.806	0.070	0.882	0.809	0.093	0.942	0.925	0.026	0.830	0.849	0.052	0.895	0.883	0.034
PoolNet-R+	2019	0.947	0.924	0.032	0.867	0.801	0.071	0.872	0.798	0.097	0.937	0.919	0.026	0.813	0.834	0.052	0.883	0.873	0.035
BASNet+	2019	0.919	0.894	0.049	0.825	0.761	0.101	0.825	0.754	0.126	0.912	0.893	0.040	0.795	0.819	0.064	0.822	0.821	0.061
DSS+ [75]	2019	0.906	0.862	0.074	0.819	0.721	0.115	0.831	0.735	0.144	0.904	0.869	0.054	0.783	0.799	0.070	0.819	0.809	0.067
PiCA-R+	2018	0.940	0.914	0.037	0.863	0.791	0.076	0.864	0.768	0.101	0.931	0.905	0.031	0.816	0.828	0.068	0.868	0.844	0.043

IV. EXPERIMENTAL RESULTS

A. Datasets and Evaluation Metrics

We used six widely-used saliency benchmark datasets in our experiments: (i) ECSSD [39] has 1,000 natural images with many semantically meaningful but complex structures; (ii) PASCAL-S [53] has 850 images generated from the PASCAL VOC2010 segmentation dataset [76], where each image has several salient objects; (iii) SOD [61] has 300 images selected from the BSDS dataset [54], where the salient objects are typically of low contrast or closely contact with the image boundary; (iv) HKU-IS [52] has 4,447 images, where most images have multiple salient objects; (v) DUT-OMRON [40] has 5,168 high-quality images, each with one or more salient objects; and (vi) DUTS [55] has a training set of 10,553 images and a testing set (denoted as DUTS-test) of 5,019 images, where the images contain various number of salient objects with large variance in scale. Among the datasets, HKU-IS, DUT-OMRON, and DUTS provide a large number of test images captured under different situations, enabling more comprehensive comparisons among different methods. Moreover, we follow the recent works on salient object detection [45], [43], [23], [25] to train our network model using the training set of DUTS [55].

Next, we used three common metrics for quantitative evaluation: F-measure (F_β), structure measure (S_m) and mean absolute error (MAE). F-measure is a balanced average precision and recall computed from the predicted maps and the ground truth images:

$$F_\beta = \frac{(1 + \beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}, \quad (6)$$

where β^2 is set as 0.3 to improve the importance of the precision, as suggested in [77], [19]. S-measure [78] computes the object-aware and region-aware structural similarity between the predicted map S and ground truth image G :

$$S_m = \alpha \times S_o(S, G) + (1 - \alpha) \times S_r(S, G), \quad (7)$$

where S_o and S_r denote the object-aware and region-aware structural similarity, respectively; α is a parameter, which balances the importance of structural similarities, and we followed [78] and set it as 0.5. Overall, a large F_β or S_m indicates a better result. MAE [36] is the average pixel-wise absolute difference between the predicted map S and the ground truth image G :

$$MAE = \frac{1}{W_S \times H_S} \sum_{x=1}^{W_S} \sum_{y=1}^{H_S} \|S(x, y) - G(x, y)\|, \quad (8)$$

where W_S and H_S are the width and height of S or G , respectively. Unlike the F_β and S_m , a small MAE indicates a better result. Finally, we used the implementation of [78], [19] to compute F_β , S_m and MAE for all results.

B. Comparison with the State-of-the-arts

We compared our method with 29 state-of-the-art methods; see the first column in Table I. Among the methods, to detect salient objects, BSCA [67] and DRFI [9] use hand-crafted features, while others employ deep neural networks to learn features. For a fair comparison, we obtained their results either by using the saliency maps provided by the authors or by producing the results using their implementations with the released training models.

Quantitative comparison. Table I summaries the quantitative results compared with the 29 state-of-the-art methods in terms of F_β , S_m and MAE on detecting salient objects in the six benchmark datasets. Our SAC-Net performs favorably against all the others for almost all the cases, regardless of whether CRF is used as a post-processing step. Especially, our method without CRF (SAC-Net) already achieves the best performance compared with all the other methods with CRF for most datasets. This result demonstrates the strong capability of our method to deal with challenging inputs; see also the visual comparison results presented in Fig. 5.

Recent deep learning methods use different kinds of backbone networks for feature extraction. For a fair comparison, we retrained these methods (PoolNet [30], BASNet [48], and PiCA [45]) by using the same backbone network (ResNet-101) as our SAC-Net. We reported the results of DSS [75] using ResNet-101 by downloading the trained model from the authors' website. These models are denoted as "XX+". Table II shows the comparison results, where our method still outperforms the very recent salient object detection methods on all the benchmark datasets. We also re-train our method by taking ResNet-50 as the backbone network, and report the results "SAC-Net (Res50)" in the last row of Table I, where our method still achieves the best performance on most of the benchmark datasets.

Visual comparison. Fig. 5 presents salient object detection results produced by various methods, including ours. From the figures, we can see that other methods (d)-(h) tend to include non-salient backgrounds or miss some salient details, while our SAC-Net is able to produce results (c) that are more consistent with the ground truth images (b). Particularly, for challenging cases, such as (i) salient objects and non-salient background with similar appearance (see 2nd and 4th rows),

TABLE III: Component analysis. Note that “SC” denotes “spatial context,” “TS” denotes “training strategy,” and “with LSTM” denotes the use of long short-term memory to aggregate the spatial context features.

	SC	TS	ECSSD			PASCAL-S			SOD			HKU-IS			DUT-OMRON			DUTS-test		
			F_β	S_m	MAE															
FPN [60]	×	SGD	0.926	0.904	0.056	0.859	0.780	0.085	0.846	0.772	0.124	0.913	0.898	0.046	0.805	0.825	0.065	0.858	0.852	0.052
SAC-Net	✓	SGD	0.949	0.928	0.036	0.878	0.805	0.072	0.874	0.806	0.099	0.938	0.923	0.030	0.828	0.849	0.055	0.888	0.879	0.038
	✓	Adam	0.951	0.931	0.031	0.879	0.806	0.070	0.882	0.809	0.093	0.942	0.925	0.026	0.830	0.849	0.052	0.895	0.883	0.034
with LSTM	✓	SGD	0.941	0.920	0.040	0.872	0.794	0.074	0.860	0.778	0.111	0.930	0.912	0.034	0.825	0.836	0.054	0.881	0.871	0.040

TABLE IV: Parameter analysis of SAC module. “ n ” is the number of attenuation factors and β is defined in Eq. (1); see Sec. III-A. Results are reported before using CRF.

n	β	HKU-IS			DUT-OMRON			DUTS-test		
		F_β	S_m	MAE	F_β	S_m	MAE	F_β	S_m	MAE
1	learnable	0.928	0.914	0.035	0.824	0.836	0.058	0.875	0.866	0.043
2	learnable	0.937	0.921	0.031	0.826	0.843	0.057	0.886	0.877	0.039
3	learnable	0.938	0.923	0.030	0.828	0.849	0.055	0.888	0.879	0.038
4	learnable	0.937	0.922	0.030	0.829	0.846	0.056	0.888	0.878	0.038
5	learnable	0.937	0.921	0.031	0.825	0.844	0.057	0.887	0.878	0.039
3	fixed (0.1)	0.936	0.921	0.031	0.825	0.846	0.056	0.887	0.877	0.039
3	fixed (0)	0.936	0.922	0.030	0.824	0.844	0.058	0.883	0.875	0.040
3	fixed (1)	0.935	0.920	0.032	0.826	0.846	0.057	0.884	0.875	0.041

(ii) small salient objects (see 2nd and 3rd rows), (iii) complex background (see 1st, 2nd, 4th, 5th, and 9th rows), and (iv) multiple objects (see 2nd, 4th, and 6th to 8th rows), our method can still predict more plausible saliency maps than the others, showing the robustness and quality of SAC-Net.

C. Evaluation on the Network Design

Component analysis. We performed an ablation study to evaluate the major components in SAC-Net. The first row of Table III shows the results from a basic model (FPN [60]) built with only the feature pyramid; see the green blocks in Fig. 2. By having the SAC modules in the network to adaptively aggregate spatial context, we can see clear improvements on all the benchmark datasets as compared with the FPN results; see the first two rows in the table.

Training strategy analysis. As mentioned in Section III-B, we adopted two different training strategies to optimize the network. The second and third rows in Table III show the comparison results, where using Adam achieves better results than using SGD. However, “Adam” took around 45 hours to train the model, while “SGD” took only around 15 hours. Hence, we adopted “SGD” to perform the following experiments to evaluate network design.

Compare with LSTM. The long short-term memory [79] (LSTM) is an efficient recurrent neural network to process sequence data by using a set of gates. The method has been extended to process 2D spatial information by some recent works on image classification [80] and saliency detection (s.t., DSCLRCN [44] and PiCA [45]). We performed another experiment by adopting the LSTMs in four principal directions with two rounds of recurrent translations to replace our recurrently attenuating translation model in the SAC module; in detail, we replaced the feature maps with colored arrows in Fig. 4 by the LSTMs in corresponding directions.

The last row in Table III presents the LSTM results. Comparing with our results in the second row, we can see

TABLE V: Architecture analysis of SAC module. Results are reported before using CRF.

Models	HKU-IS			DUT-OMRON			DUTS-test		
	F_β	S_m	MAE	F_β	S_m	MAE	F_β	S_m	MAE
one-round	0.933	0.920	0.032	0.824	0.845	0.056	0.883	0.875	0.040
three-round	0.937	0.922	0.031	0.828	0.848	0.055	0.886	0.879	0.039
w/o left-right	0.935	0.920	0.031	0.823	0.843	0.057	0.885	0.876	0.039
w/o up-down	0.936	0.921	0.031	0.824	0.843	0.056	0.886	0.877	0.039
w/o attention	0.934	0.919	0.032	0.824	0.844	0.055	0.884	0.876	0.039
Ours	0.938	0.923	0.030	0.828	0.849	0.055	0.888	0.879	0.038

that our method performs better for F_β , S_m and MAE on all the benchmark data. We think the reason is that due to the limitation of the gate functions in LSTM [81], context features can only propagate over a short distance, thus limiting the dispersal of local context features in the spatial domain. On the other hand, the time complexity of computing LSTMs on 2D feature maps is very high. “with LSTM” took around 213 hours to train the model, while our method took only around 15 hours, which is more than 14 times faster.

Parameter analysis. To build our network, we empirically determine the value of n , which affects the number of attenuation factors and the number of feature channels in each aggregated feature map ($\lfloor \frac{256}{n} \rfloor$); see Fig. 4. In general, a large n allows the network to consider more variety of attenuation factors but each feature map would capture less information in return, since we keep the overall memory consumption to be manageable. Another parameter in our network is β , where we automatically learn its value for regulating the magnitude of the negative part in Eq. (1).

We evaluated our network on the three largest datasets (HKU-IS, DUT-OMRON, and DUTS-test) using different n and learnable/fixed β . The results shown in Table IV reveal that when we aggregate the image context using two different attenuation factors ($n=2$), we achieve better results than using only one single long-range aggregation ($n=1$). The results further improve with larger n and roughly stabilizes when n reaches three, so we set $n=3$. On the other hand, comparing the results on the 3rd and last three rows (all with $n=3$) in table, we can see that automatically learning and adjusting β gives better results than using a fixed β ($\beta = 0.1$ or 0), or linearly aggregating the spatial features ($\beta = 1$).

Architecture analysis. To evaluate the effectiveness of our network design, we construct several variant models of our network. As shown in the Table V, first, we replace the two-round recurrent translations in our SAC module by one-round and three-round. The results show that our method with two-round recurrent translations achieves the best performance. Then, we build two modules, i.e., “w/o left-right” and “w/o

TABLE VI: Comparing our method (SAC-Net) with other works on spatial context using ResNet-101 as the backbone network. Results are reported before using CRF.

Dataset	ECSSD [39]			PASCAL-S [53]			SOD [61]			HKU-IS [52]			DUT-OMRON [40]			DUTS-test [55]		
	F_β	S_m	MAE	F_β	S_m	MAE	F_β	S_m	MAE	F_β	S_m	MAE	F_β	S_m	MAE	F_β	S_m	MAE
SAC-Net (ours)	0.951	0.931	0.031	0.879	0.806	0.070	0.882	0.809	0.093	0.942	0.925	0.026	0.830	0.849	0.052	0.895	0.883	0.034
DSC [82], [83]	0.948	0.929	0.036	0.877	0.801	0.072	0.872	0.801	0.100	0.935	0.920	0.031	0.830	0.847	0.053	0.886	0.878	0.038
DeepLabv3+ [84]	0.947	0.925	0.037	0.878	0.797	0.071	0.862	0.788	0.102	0.934	0.915	0.032	0.824	0.836	0.053	0.885	0.872	0.038
PSPNet [85]	0.940	0.917	0.042	0.877	0.795	0.071	0.860	0.777	0.111	0.927	0.911	0.036	0.819	0.829	0.056	0.881	0.869	0.040
PSANet [86]	0.940	0.917	0.042	0.873	0.796	0.073	0.858	0.778	0.112	0.928	0.912	0.036	0.816	0.831	0.056	0.879	0.869	0.041
DeepLabv3 [87]	0.939	0.917	0.042	0.873	0.793	0.073	0.862	0.775	0.111	0.926	0.909	0.037	0.821	0.827	0.056	0.877	0.865	0.042
Non-local Network [34]	0.936	0.915	0.044	0.874	0.795	0.072	0.858	0.776	0.112	0.924	0.906	0.037	0.809	0.826	0.059	0.873	0.865	0.043
DeepLab [88]	0.934	0.914	0.045	0.873	0.797	0.073	0.850	0.774	0.114	0.923	0.907	0.038	0.803	0.822	0.059	0.871	0.863	0.043

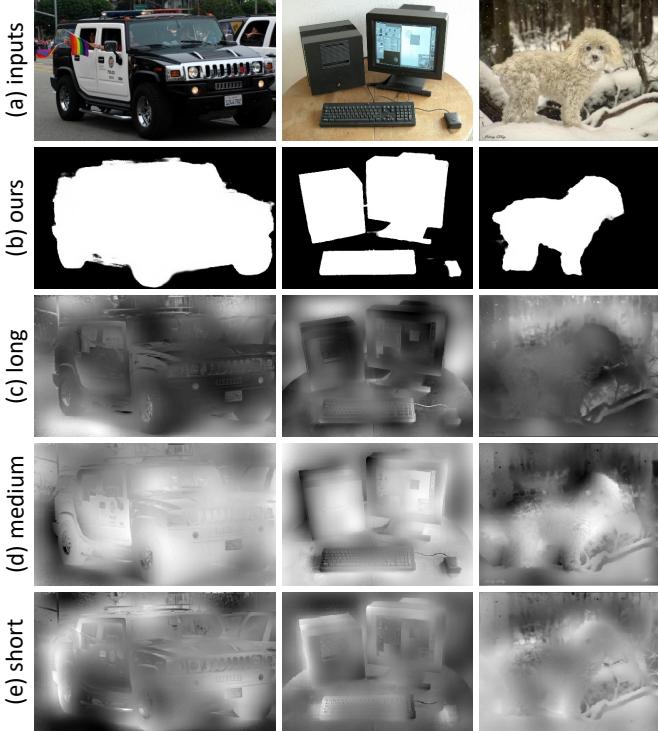


Fig. 6: Attention weights learned for different spatial ranges, where the brightness indicates the magnitude of the learned attention weights.

up-down”, by removing the recurrent translations in the left and right or up and down directions, which leads to the worse results. Finally, we remove the attention mechanism in the SAC module to build “w/o attention”. Results show that our network design achieves the best performance.

Attention weight visualization. Figs. 1 & 6 visualize the learned attention weights for integrating the spatial context features. The long-range context (c) helps to locate the global background regions; the medium-range context (d) helps to identify the image regions of objects; and short-range context (e) helps to locate the boundary between salient and non-salient regions. Moreover, our attention mechanism selectively aggregates various spatial context and allows the context features to be implicitly dispersed over arbitrary spatial ranges.

Time performance. Our network is fast, since it has a fully convolutional architecture and employs an efficient recurrent translation module. We tested our network on a single GPU

TABLE VII: Time performance analysis. “FPS” stands for “frames per second.”

Method	Ours	PiCANet [45]	DGRL [43]	R ³ Net [18]
FPS	11	7	8	4
Method	SRM [63]	Amulet[24]	NLDF [13]	DSS [19]
FPS	14	16	12	12

TABLE VIII: Comparing with state-of-the-art methods on shadow detection. All the deep-learning methods are trained on the SBU training set and tested on the SBU testing set.

Method	BER
shadow detection	BDRAR [89]
	DSC [82], [83]
	scGAN [90]
	stacked-CNN [91]
	patched-CNN [92]
saliency detection	Unary-Pairwise [93]
	SAC-Net (ours)
	R ³ Net [18]
	PiCANet [45]
	RADF [20]
	SRM [63]
	RAS [17]
Amulet [24]	15.13

(TITAN Xp) using input images of size 400×400 . It takes around 0.090 seconds on average to test one image. If we remove the SAC modules from our network, it still needs 0.087 seconds to process one image, which proves the efficiency of the proposed SAC module. Moreover, we compare the time performance of our SAC-Net with other methods for salient object detection. Table VII shows the results, where our method has comparable time performance with other methods that have worse detection accuracy than ours.

D. Shadow Detection

Our SAC model has the potential to be applied to other vision tasks. Here, we take the shadow detection as an example. We re-train our network as well as other salient object detection methods on the training set of SBU [91], which is a widely used dataset for shadow detection, and test them on the testing set of SBU. Moreover, we use the common metric BER for the quantitative comparisons among different shadow detectors. Table VIII reports the results, where our SAC-Net achieves the best performance among the methods designed

for salient object detection and also outperforms most of the shadow detection methods.

V. DISCUSSION

There has been a lot of works on exploiting spatial context in deep CNNs for image analysis. Dilated convolution [88], [94] takes context from larger regions by inserting holes into the convolution kernels, but the context information in use still has a fixed range in a local region. ASPP [84], [87] and PSPNet [85] adopt multiple convolution kernels with different dilated rates or multiple pooling operations with different scales to aggregate spatial context using different region sizes; however, their designed kernel or pooling sizes are fixed, less flexible, and not adaptable to different inputs. DSC [82], [83] adopts the attention weights to indicate the importance of context features aggregated from different directions, but it only obtains the global context with a fixed influence range over the spatial domain. The non-local network [34] computes correlations between every pixel pair on the feature map to encode the global image semantics, but this method ignores the spatial relationship between pixels in the aggregation; for salient object detection, features of opposite semantics may, however, be important; see Fig. 1. PSANet [86] adaptively learns attention weights for each pixel to aggregate the information from different positions; however, it is unable to capture the context on lower-level feature maps in high resolutions due to the huge time and memory overhead. Compared to these methods, our SAC-Net explores and adaptively aggregates context features implicitly with variable influence ranges; it is flexible, fast, and computationally friendly for efficient salient object detection.

We performed an experiment by training these methods on the DUTS training set for salient object detection. For a fair comparison, we adopted ResNet-101 as the backbone network for all the methods. Table VI reports the results, where our method still achieves the best performance on all the benchmark datasets, which proves the effectiveness of the designed SAC module.

Lastly, we also analyzed the failure cases, for which we found to be highly challenging. For instance, our method may fail for (i) multiple salient objects in very different scales (see Fig. 7 (top)), where the network may regard the small objects as non-salient background; (ii) dark salient objects (see Fig. 7 (middle)), where there are insufficient context to determine whether the regions are salient or not; and (iii) salient objects over a complex background (see Fig. 7 (bottom)), where high-level scene knowledge is required to understand the image.

VI. CONCLUSION

This paper presents a novel saliency detection network based on the spatial attenuation context. Our key idea is to recurrently propagate and aggregate image context with different attenuation factors and to integrate the aggregated features using weights learnt from an attention mechanism. Using our model, local image context can adaptively propagate over different ranges, and we can leverage the complementary advantages of these context to improve the saliency detection

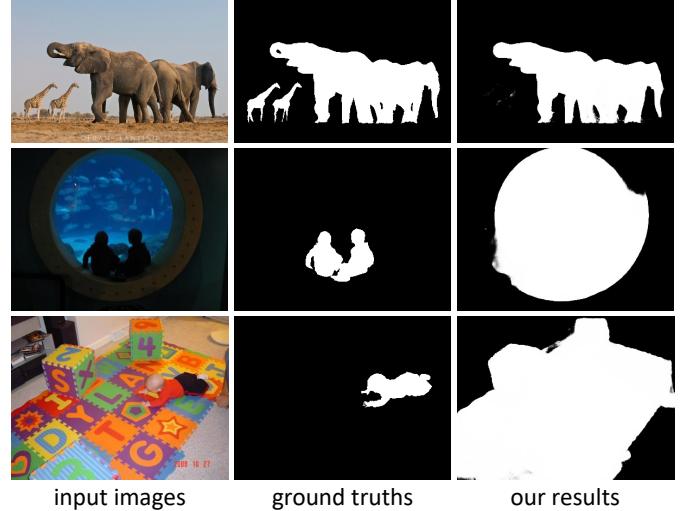


Fig. 7: Three typical failure cases.

quality. In the end, we evaluated our method on six common benchmark datasets and compared it extensively with 29 state-of-the-art methods. Experimental results clearly show that our method performs favorably over all the others, both visually and quantitatively. In the future, we plan to explore the potential of our SAC module design for instance-level salient object detection and enhance its capability for detecting salient objects in videos.

REFERENCES

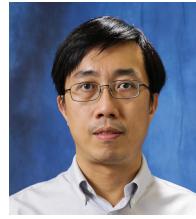
- [1] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *CVPR*, 2015, pp. 3395–3402.
- [2] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," in *ICML*, 2015, pp. 597–606.
- [3] H. Hadizadeh and I. V. Bajic, "Saliency-aware video compression," *IEEE Transactions on Image Processing*, vol. 23, no. 1, pp. 19–33, 2014.
- [4] H. Zhao, X. Mao, X. Jin, J. Shen, F. Wei, and J. Feng, "Real-time saliency-aware video abstraction," *The Visual Computer*, vol. 25, no. 11, pp. 973–984, 2009.
- [5] M.-M. Cheng, F.-L. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Repfinder: Finding approximately repeated scene elements for image editing," in *ACM Trans. on Graphics (SIGGRAPH)*, vol. 29, no. 4, ACM, 2010, p. 83.
- [6] L. Zhu, X. Hu, C.-W. Fu, J. Qin, and P.-A. Heng, "Saliency-aware texture smoothing," *IEEE Transactions on Visualization and Computer Graphics*, 2018.
- [7] H. Zhang, J. Zhang, and P. Koniusz, "Few-shot learning via saliency-guided hallucination of samples," in *CVPR*, 2019, pp. 2770–2779.
- [8] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2015.
- [9] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *CVPR*, 2013, pp. 2083–2090.
- [10] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 353–367, 2011.
- [11] C. Deng, X. Yang, F. Nie, and D. Tao, "Saliency detection via a multiple self-weighted graph-based manifold ranking," *IEEE Transactions on Multimedia*, 2019.
- [12] X. Li, L. Zhao, L. Wei, M.-H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "DeepSaliency: Multi-task deep neural network model for salient object detection," *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3919–3930, 2016.
- [13] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," in *CVPR*, 2017, pp. 6609–6617.

- [14] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *ECCV*, 2016, pp. 825–841.
- [15] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *ICCV*, 2017, pp. 212–221.
- [16] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *CVPR*, 2015, pp. 1265–1274.
- [17] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *ECCV*, 2018.
- [18] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, and P.-A. Heng, "R³Net: Recurrent residual refinement network for saliency detection," in *IJCAI*, 2018, pp. 684–690.
- [19] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr, "Deeply supervised salient object detection with short connections," in *CVPR*, 2017, pp. 3203–3212.
- [20] X. Hu, L. Zhu, J. Qin, C.-W. Fu, and P.-A. Heng, "Recurrently aggregating deep features for salient object detection," in *AAAI*, 2018, pp. 6943–6950.
- [21] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *CVPR*, 2016, pp. 478–487.
- [22] X. Li, F. Yang, H. Cheng, W. Liu, and D. Shen, "Contour knowledge transfer for salient object detection," in *ECCV*, 2018.
- [23] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang, "A bi-directional message passing model for salient object detection," in *CVPR*, 2018, pp. 1741–1750.
- [24] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *ICCV*, 2017, pp. 202–211.
- [25] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *CVPR*, 2018, pp. 714–722.
- [26] P. Zhang, W. Liu, H. Lu, and C. Shen, "Salient object detection with lossless feature reflection and weighted structural loss," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 3048–3060, 2019.
- [27] W. Wang, S. Zhao, J. Shen, S. C. Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *CVPR*, 2019, pp. 1448–1457.
- [28] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *CVPR*, 2019, pp. 3085–3094.
- [29] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *CVPR*, 2019, pp. 3907–3916.
- [30] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *CVPR*, 2019, pp. 3917–3926.
- [31] W. Wang, J. Shen, M.-M. Cheng, and L. Shao, "An iterative and cooperative top-down and bottom-up inference network for salient object detection," in *CVPR*, 2019, pp. 5968–5977.
- [32] R. Wu, M. Feng, W. Guan, D. Wang, H. Lu, and E. Ding, "A mutual learning method for salient object detection with intertwined multi-supervision," in *CVPR*, 2019, pp. 8150–8159.
- [33] G. Li and Y. Yu, "Contrast-oriented deep neural networks for salient object detection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 12, pp. 6038–6051, 2018.
- [34] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *CVPR*, 2018, pp. 7794–7803.
- [35] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1915–1926, 2012.
- [36] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *CVPR*, 2012, pp. 733–740.
- [37] A. Borji and L. Itti, "Exploiting local and global patch rarities for saliency detection," in *CVPR*, 2012, pp. 478–485.
- [38] V. Mahadevan and N. Vasconcelos, "Biologically inspired object tracking using center-surround saliency mechanisms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 541–554, 2013.
- [39] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *CVPR*, 2013, pp. 1155–1162.
- [40] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *CVPR*, 2013, pp. 3166–3173.
- [41] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *NIPS*, 2007, pp. 545–552.
- [42] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Saliency object detection: A benchmark," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5706–5722, 2015.
- [43] 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 *CVPR*, 2018, pp. 3127–3135.
- [44] N. Liu and J. Han, "A deep spatial contextual long-term recurrent convolutional network for saliency detection," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3264–3274, 2018.
- [45] N. Liu, J. Han, and M.-H. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," in *CVPR*, 2018, pp. 3089–3098.
- [46] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *CVPR*, 2019, pp. 1623–1632.
- [47] L. Zhang, J. Zhang, Z. Lin, H. Lu, and Y. He, "CapSal: Leveraging captioning to boost semantics for salient object detection," in *CVPR*, 2019, pp. 6024–6033.
- [48] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *CVPR*, 2019, pp. 7479–7489.
- [49] K. Fu, Q. Zhao, I. Y.-H. Gu, and J. Yang, "Deepside: A general deep framework for salient object detection," *Neurocomputing*, vol. 356, pp. 69–82, 2019.
- [50] G. Li, Y. Xie, L. Lin, and Y. Yu, "Instance-level salient object segmentation," in *CVPR*, 2017, pp. 2386–2395.
- [51] L. Zhu, J. Chen, X. Hu, C.-W. Fu, X. Xu, J. Qin, and P.-A. Heng, "Aggregating attentional dilated features for salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.
- [52] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *CVPR*, 2015, pp. 5455–5463.
- [53] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *CVPR*, 2014, pp. 280–287.
- [54] 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 *ICCV*, 2001, pp. 416–423.
- [55] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *CVPR*, 2017, pp. 136–145.
- [56] K. Song, H. Yang, and Z. Yin, "Multi-scale attention deep neural network for fast accurate object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 10, pp. 2972–2985, 2019.
- [57] Y. Peng, Y. Zhao, and J. Zhang, "Two-stream collaborative learning with spatial-temporal attention for video classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 3, pp. 773–786, 2019.
- [58] X. He, Y. Peng, and J. Zhao, "Which and how many regions to gaze: Focus discriminative regions for fine-grained visual categorization," *International Journal of Computer Vision*, vol. 127, no. 9, pp. 1235–1255, 2019.
- [59] Y. Peng, X. He, and J. Zhao, "Object-part attention model for fine-grained image classification," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1487–1500, 2018.
- [60] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017, pp. 2117–2125.
- [61] V. Movahedi and J. H. Elder, "Design and perceptual validation of performance measures for salient object segmentation," in *CVPRW*, 2010, pp. 49–56.
- [62] W. Wang, J. Shen, X. Dong, and A. Borji, "Salient object detection driven by fixation prediction," in *CVPR*, 2018, pp. 1711–1720.
- [63] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, "A stagewise refinement model for detecting salient objects in images," in *CVPR*, 2017, pp. 4019–4028.
- [64] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *CVPR*, 2016, pp. 678–686.
- [65] G. Lee, Y.-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *CVPR*, 2016, pp. 660–668.
- [66] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *CVPR*, 2015, pp. 3183–3192.
- [67] Y. Qin, H. Lu, Y. Xu, and H. Wang, "Saliency detection via cellular automata," in *CVPR*, 2015, pp. 110–119.
- [68] Y. Wu and K. He, "Group normalization," in *ECCV*, 2018, pp. 3–19.
- [69] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.
- [70] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

- [71] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.
- [72] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.
- [73] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
- [74] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *NIPS*, 2011, pp. 109–117.
- [75] Q. Hou, M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 4, p. 815, 2019.
- [76] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.
- [77] R. Achanta, S. Hemami, F. Estrada, and S. Sussstrunk, "Frequency-tuned salient region detection," in *CVPR*, 2009, pp. 1597–1604.
- [78] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *ICCV*, 2017, pp. 4548–4557.
- [79] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [80] F. Visin, K. Kastner, K. Cho, M. Matteucci, A. Courville, and Y. Bengio, "ReNet: A recurrent neural network based alternative to convolutional networks," *arXiv preprint arXiv:1505.00393*, 2015.
- [81] Q. V. Le, N. Jaitly, and G. E. Hinton, "A simple way to initialize recurrent networks of rectified linear units," *arXiv preprint arXiv:1504.00941*, 2015.
- [82] X. Hu, L. Zhu, C.-W. Fu, J. Qin, and P.-A. Heng, "Direction-aware spatial context features for shadow detection," in *CVPR*, 2018, pp. 7454–7462.
- [83] X. Hu, C.-W. Fu, L. Zhu, J. Qin, and P.-A. Heng, "Direction-aware spatial context features for shadow detection and removal," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
- [84] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.
- [85] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *CVPR*, 2017, pp. 2881–2890.
- [86] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. Change Loy, D. Lin, and J. Jia, "PSANet: Point-wise spatial attention network for scene parsing," in *ECCV*, 2018, pp. 267–283.
- [87] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.
- [88] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *ICLR*, 2015.
- [89] L. Zhu, Z. Deng, X. Hu, C.-W. Fu, X. Xu, J. Qin, and P.-A. Heng, "Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection," in *ECCV*, 2018, pp. 121–136.
- [90] V. Nguyen, T. F. Y. Vicente, M. Zhao, M. Hoai, and D. Samaras, "Shadow detection with conditional generative adversarial networks," in *ICCV*, 2017, pp. 4510–4518.
- [91] T. F. Y. Vicente, L. Hou, C.-P. Yu, M. Hoai, and D. Samaras, "Large-scale training of shadow detectors with noisily-annotated shadow examples," in *ECCV*, 2016, pp. 816–832.
- [92] S. Hosseinzadeh, M. Shakeri, and H. Zhang, "Fast shadow detection from a single image using a patched convolutional neural network," in *International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 3124–3129.
- [93] R. Guo, Q. Dai, and D. Hoiem, "Single-image shadow detection and removal using paired regions," in *CVPR*, 2011, pp. 2033–2040.
- [94] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *ICLR*, 2016.



Xiaowei Hu received his B.Eng. degree in Computer Science and Technology from South China University of Technology, China, in 2016. He is currently working toward the Ph.D. degree with the Department of Computer Science and Engineering, The Chinese University of Hong Kong. His research interests include computer vision, deep learning, and low-level vision.



Chi-Wing Fu is currently an associate professor in the Chinese University of Hong Kong. He served as the co-chair of SIGGRAPH ASIA 2016's Technical Brief and Poster program, associate editor of Computer Graphics Forum, and panel member in SIGGRAPH 2019 Doctoral Consortium, as well as program committee members in various research conferences, including SIGGRAPH Asia Technical Brief, SIGGRAPH Asia Emerging tech., IEEE visualization, CVPR, IEEE VR, VRST, Pacific Graphics, GMP, etc. His recent research interests include computation fabrication, 3D computer vision, user interaction, and data visualization.



Lei Zhu received his Ph.D. degree in the Department of Computer Science and Engineering from the Chinese University of Hong Kong in 2017. He is working as a postdoctoral fellow at the Chinese University of Hong Kong. His research interests include computer graphics, computer vision, medical image processing, and deep learning.



Tianyu Wang received his B.Eng. degree in Computer Science and Technology from Dalian University of Technology, China, in 2018. He is currently working as a research assistant at the Chinese University of Hong Kong. His research interests include computer vision, image processing, computational photography, low-level vision, and deep learning.



Pheng-Ann Heng received his B.Sc. (Computer Science) from the National University of Singapore in 1985. He received his M.Sc. (Computer Science), M. Art (Applied Math) and Ph.D. (Computer Science) all from the Indiana University in 1987, 1988, 1992 respectively. He is a professor at the Department of Computer Science and Engineering at The Chinese University of Hong Kong. He has served as the Department Chairman from 2014 to 2017 and as the Head of Graduate Division from 2005 to 2008 and then again from 2011 to 2016. He has served as the Director of Virtual Reality, Visualization and Imaging Research Center at CUHK since 1999. He has served as the Director of Center for Human-Computer Interaction at Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences since 2006. He has been appointed by China Ministry of Education as a Cheung Kong Scholar Chair Professor in 2007. His research interests include AI and VR for medical applications, surgical simulation, visualization, graphics, and human-computer interaction.