Pyramid Spatial Context Features for Salient Object Detection

HUI LI
School of Computer Science and Engineering, Huizhou University, Huizhou 516007, China
e-mail: zjhz_lihui@163.com

This work was supported in part by the Educational Science Planning Project of Guangdong Province “Research on Innovative Talent Fostering Pattern with Deep Converging of School and Industries under New Engineering Disciplines” under Grant 2018GXJK183, and in part by the Indigenous Innovation’s Capability Development Program of Huizhou University under Grant hzu202010.

ABSTRACT Salient object detection aims to obtain the most attractive objects from the input images, which serves as a pre-processing step for many image processing tasks. This paper presents a novel deep neural network design for salient object detection by formulating a pyramid spatial context module, PSC module for short, to capture the spatial context information at multiple scales. To achieve this, we first adopt convolutional operations with different dilated rates to generate the feature maps with different respective fields, and then use the two-round recurrent translations to explore multiple types of spatial context features on these feature maps. By further inserting this module in a deep network, namely PSCNet, we are able to optimize the network in an end-to-end manner for salient object detection. We evaluate the proposed method on six public benchmark datasets by comparing it with 25 salient object detection methods. The experimental results demonstrate that our PSCNet performs favorably against all the other methods.

INDEX TERMS Salient object detection, spatial context, pyramid features, deep neural network.

I. INTRODUCTION Saliency detection is one of useful pre-processing steps for lots of image processing tasks, including video compression [1], video abstraction [2], image editing [3], texture smoothing [4], object detection [5], [6], and few-shot learning [7]. It aims to find the most obvious objects from the input image and this task has been widely studied in the past years.

Early works adopt low-level vision cues to identify the salient objects, where these visual cues include contrast, color, and texture [8]–[10]. However, the low-level cues lack high-level semantic information, which is important to distinguish the salient objects from the background. Hence, the recent works [11]–[15] leverage deep convolutional neural networks (CNNs) to learn to detect salient objects while more and more works [16]–[33] use both highly semantic features from the deep CNN layers and low-level features from shallow CNN layers for saliency detection.

Moreover, very recent works [35], [36] explore the non-local features for saliency detection by further reasoning the spatial context from the whole feature map. These context features are obtained by propagating the information from one pixel to the adjacent pixel and this process is repeated over the whole feature map. Hence, for each pixel, it can obtain the global context information from any position of the feature map. However, this process is performed on the feature map generated by a single convolutional operation and we can only obtain a single type of spatial context feature, which is difficult to capture the complex structures of salient objects and noisy background. Also, their results tend to lose some salient details or include background noises, as shown in Figure 1.

In this paper, we present to formulate a pyramid spatial context module, PSC module for short, to generate multiple types of spatial context features, which are able to capture the spatial context information at different scales, thus improving the performance of salient object detection by analyzing the fine-grain and coarse-grain characteristics of the salient objects. Our module first performs the convolutional operations with different dilated rates to generate the convolutional features with multiple respective fields, and then leverages two-round recurrent translations to each convolutional feature map individually to obtain multiple types of spatial context features. The spatial context features aggregated from the feature map with the small respective field tend to explore the fine-grain characteristics of the images, which helps to...
enhance the detail structures of the salient objects, while the spatial context features aggregated from the feature map with the large respective field tend to obtain the coarse-grain characteristics of the images, which helps to distinguish the salient objects from the background. By jointly using different types of spatial context information, we are able to capture the image features at multiple levels and produce more accurate detection results, as shown in Figure 1(b). Next, we insert the PSC module in a deep convolutional neural network to build the PSCNet for salient object detection. Finally, we use six benchmark datasets to evaluate the effectiveness of our method by comparing it with 25 state-of-the-art methods for salient object detection. Experimental results show that we achieve the best performance on all the benchmark datasets. We summarize the contributions of this paper as follows:

- First, we formulate a new module, named as the pyramid spatial context (PSC) module, to capture the different kinds of spatial context features on the convolutional feature maps with multiple respective fields.
- Second, we build a novel deep convolutional neural network architecture, PSCNet, by adopting the PSC module to aggregate multi-level spatial context features and train the network in an end-to-end manner for salient object detection.
- Third, we compare our PSCNet with other 25 methods for salient object detection on six benchmark datasets by using three evaluation metrics. Experimental results show that our method outperforms other methods on all the benchmark datasets.

II. RELATED WORK

Salient object detection, especially based on the single input image, has been widely studied in the past years. Early methods adopt the hand-crafted features to detect salient objects and these hand-crafted features include color [37], [38], texture [39], [40], image contrast [9], [41], and other types of low-level visual cues [42]; please see [43] for an exhaustive survey. However, these hand-crafted features have limited feature expression ability, leading to a failure to find the salient objects in the complex environment.

With the development of deep convolutional neural networks (CNNs), lots of works leverage the highly semantic features learned from CNNs automatically for salient object detection. Recently, more and more works [16]–[33], [44] try to jointly use the low-level detail information and high-level semantic information at the different layers of CNN and achieve more accurate saliency detection results. For example, Li and Yu [20] used multi-layer features with different resolutions to find the semantic characters and the visual contrast of salient objects. Hou et al. [18] developed short connections to merge the features at multiple layers for salient object predictions. Zhang et al. [23] formulated a boundary preserved refinement strategy and integrated multi-layer feature maps to predict the saliency details. Hu et al. [19] recurrently aggregated deep features from a deep network to leverage the complementary information among multiple CNN layers. Deng et al. [17] explored the residual learning strategy to alternatively enhance the high-level features at deep layers and low-level features at shallow layers. Zhang et al. [22] designed a bi-directional message-passing model to select the useful features by a gate function. Zhang et al. [24] utilized an attention guided network to progressively and selectively integrate multi-level features. Wang et al. [26] developed a pyramid attention structure to better predict the salient objects. Zhu et al. [33] explored the local and global information of CNN by learning the attentional dilated features. Feng et al. [45] formulated an attentional feedback network for salient object detection by exploring the salient boundaries. Zhao and Wu [27] introduced the dilated convolution and channel-wise and spatial-wise attentions to use multi-scale features. Wu et al. [28] discarded the shallow-layer feature maps at shallow layers for acceleration. Wang et al. [30] iteratively integrated feature maps in the top-down and bottom-up manner. Qin et al. [34] developed a boundary-aware salient object detection network for salient object detection. Wu et al. [31] leveraged the advantages of multi-task learning to formulate foreground contour detection and edge detection tasks as the supervisions to enhance the network ability to predict more accurate salient objects.

Very recently, Liu and Han [35], [36] incorporated the global context feature to detect salient objects by adopting the spatial long short-term memory model. Zhao [15] adopts fully connected operations and Chen et al. [46] uses global average pooling to obtain global information. The fully connected operations and spatial long short-term memory model are time-consuming and global average pooling can only obtain a single feature for all the pixel on the spatial domain of feature maps. Beyond these works that aggregate a single type of context features, we present to generate multiple levels of spatial context features to capture both fine-grain and coarse-grain characteristics of the salient objects.
Moreover, please also refer [47] for a review of visual saliency detection with comprehensive information, [48]–[52] for salient object detection in RGBD images, [53] for salient object detection in optical remote sensing images, [54] for video salient object detection, and [55] for salient object detection in stereoscopic images.

III. METHODOLOGY
A. NETWORK OVERVIEW
Figure 2 illustrates the overall architecture of our pyramid spatial context network (PSCNet), which uses a single image as the input and outputs a predicted saliency map in an end-to-end manner. In the beginning, we adopt a deep convolutional neural network to generate the convolutional feature maps with different resolutions, and the feature maps with large resolutions at shallow layers help to discover the detail structures while the feature maps with small resolutions at deep layers help to generate the highly semantic features to distinguish the salient objects from the background. During the implementation, we take the DeepLab [56] as the feature extraction network, which keeps the resolutions of the last two layers unchanged and uses the convolutional operations with the dilated rates to enlarge the receptive fields. After obtaining the convolutional features, we leverage the proposed pyramid spatial context (PSC) module to aggregate multi-level context features, and finally use a 1×1 convolution to reduce the number of feature channels and obtain $F_o$. Note that we adopt batch normalization [57] and the ReLU [58] non-linear transformation after each convolutional operation. In the end, we predict the saliency map from $F_o$ as the network output.

B. SPATIAL CONTEXT FEATURES
To aggregate the spatial context features from the 2D feature maps, some works [70]–[72] present to perform the two-round data translations on the spatial domain of the feature maps to propagate the context information pixel by pixel. As shown in Figure 3, given the input feature map, we adopt the data translations of four directions (up, down, left, right), and for each pixel $x_{ij}$ on the feature map (as the green point shown in Figure 3(a)), it can obtain the information from its four adjacent pixels:

$$x^\text{up}_{ij} = \max(\omega^{\text{up}}x_{i+1,j} + x_{ij}, 0)$$  \hspace{1cm} (1)
$$x^\text{down}_{ij} = \max(\omega^{\text{down}}x_{i-1,j} + x_{ij}, 0)$$  \hspace{1cm} (2)
$$x^\text{left}_{ij} = \max(\omega^{\text{left}}x_{i,j+1} + x_{ij}, 0)$$  \hspace{1cm} (3)
$$x^\text{right}_{ij} = \max(\omega^{\text{right}}x_{i,j-1} + x_{ij}, 0)$$  \hspace{1cm} (4)

where $\omega^{\text{up}}$, $\omega^{\text{down}}$, $\omega^{\text{left}}$, and $\omega^{\text{right}}$ are the parameters that can be learned automatically, and we set them as one at the beginning of the training process. The learned $\omega$ usually belongs to $[0,1]$. These recurrent translations at four directions are performed individually over the whole feature maps. After we combine the results of four directions, for each pixel, it can obtain the spatial context information from its row and column over the whole feature maps; see Figure 3(b).

Next, we take the feature map $F_m$ in Figure 3(b) as the input and perform the second-round recurrent translation on
the four directions, and combine the results as the output. At this time, each pixel can obtain the context information from the whole feature map, as shown in Figure 3(c).

### C. Pyramid Spatial Context Features

The spatial context features obtained in Section III-B are largely affected by the adjacent pixel when the parameter $\omega$ is less than one. To generate multiple types of spatial context features at multiple scales with larger receptive fields, motivated by ASPP [56], we first adopt the convolution with different dilated rates to produce a set of feature maps. As shown in Figure 2, after we obtain the feature map extracted by the feature extraction network, we apply (i) a convolutional operation with the kernel size of one, (ii) a convolutional operation with the kernel size of three and dilation rate of six, (iii) a convolutional operation with the kernel size of three and dilation rate of 12, and (iv) a convolutional operation with the kernel size of three and dilation rate of 18. Also, we perform the batch normalization and ReLU non-linear operation at each convolutional operation. Hence, we can obtain four feature maps with different respective fields.

Next, we adopt the two-round data translations at four directions for each feature map to obtain the spatial context features at different scales. Further, we concatenate the spatial context features obtained from four feature maps and also combine them with the original convectional feature; see the orange block in Figure 2. Since the scales of the spatial context features are different, we name the aggregated features as pyramid spatial context features and name this module as the pyramid spatial context module. Finally, we apply a convolutional operation with the kernel size of one to merge the concatenated features and obtain the feature map $F_o$ with fewer feature channels, and the feature map $F_o$ is used to predict the saliency map as the output of the network.

### D. Training and Testing Strategies

1) Loss Function

Cross-entropy loss was used to optimize our PSCNet:

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

where $g$ and $p$ denote the ground truth image and the predicted saliency map, respectively; $g_{i,j}$ is zero or one while $p_{i,j}$ belongs to $[0, 1]$.

2) Training Parameters

The feature extraction part in our PSCNet (the orange blocks in Figure 2) was initialized by the weights trained on ImageNet [73], and other parts of our network were initialized by random noise. Next, we adopted Adam [74] to optimize the network by setting the first momentum value as 0.9, the second momentum value as 0.999, and the weight decay of 0.0005. We set the learning rate as 0.00005, reduce it to 0.000005 after 160k training iterations, and terminate the training process after 250k iterations. At the same time, we horizontally flipped the images for data argumentation and used the mini-batch size of one. Our method uses around six days to train the whole network a single TITAN X (Pascal) GPU.

3) Inference

We use the predicted saliency map of our PSCNet as the output, and upsample it to the size of the original input image. Our method adopts 0.159s to process an input image on average on a single TITAN X (Pascal) GPU.

### E. Discussion

To obtain various types of spatial context, firstly, we adopt the convolutions with different dilated rates to obtain the convolutional features. When the dilated rate is zero, we mainly focus on the local regions, which helps to discover the fine detail information. When the dilated rate increases, we can obtain the convolutional features with large respective fields, which helps to discover the whole saliency region. Second, we take the convolutional feature maps with different receptive fields as the inputs to the two-round data translations to obtain the global spatial context. The spatial context aggregated from the feature maps with a small receptive field reflects the fine-grain characteristic of the salient objects, since each pixel of the feature map only contains the information of itself and only propagates its information to other pixels. In contrast, the spatial context aggregated from the feature maps with a large receptive field reflects the coarse-grain characteristic of the salient objects, since each pixel of the feature map contains the information of a local region and propagates the information of such a local region to other pixels. Hence, by combining the multiple types of spatial context features, we can capture the spatial context information at different scales and improve the performance.

---

**TABLE 1.** Component analysis. Note that “+SC” indicates we only use spatial context in our network and “PSCNet” denotes we use the proposed pyramid spatial context to detect salient objects.

|                | ECSSD | PASCAL-S | SOD | HKU-IS | DUT-OMRON | DUTS-test |
|----------------|-------|----------|-----|--------|-----------|-----------|
|                | $F_o$ | $S_m$    | MAE | $F_o$  | $S_m$    | MAE       | $F_o$  | $S_m$    | MAE       | $F_o$  | $S_m$    | MAE       |
| Deeplab        | 0.934 | 0.914    | 0.045 | 0.873  | 0.797    | 0.073     | 0.850  | 0.774    | 0.114     | 0.923  | 0.907    | 0.038     |
| +SC            | 0.938 | 0.917    | 0.043 | **0.876** | 0.794    | 0.073     | 0.862  | 0.779    | 0.111     | 0.927  | 0.910    | 0.037     |
| one-round      | 0.940 | 0.918    | 0.040 | 0.863  | 0.795    | 0.075     | 0.862  | 0.782    | 0.104     | 0.925  | 0.902    | 0.040     |
| three-round    | 0.942 | 0.925    | 0.035 | 0.868  | 0.799    | 0.072     | 0.869  | 0.790    | 0.100     | 0.929  | 0.914    | 0.031     |
| PSCNet (ours)  | **0.944** | **0.924** | **0.031** | 0.864  | **0.805** | **0.071** | 0.870  | **0.802** | **0.095** | 0.932  | **0.916** | **0.026** |

VOLUME 8, 2020 88521
TABLE 2. Comparison results between our method (PSCNet) and other methods using the $F_\beta$, $S_m$, and MAE metrics. ‘-’ indicates results that are not publicly available; We don’t use any post-processing method, e.g., CRF.

| Dataset            | Metric | Year   | $F_\beta$ | $S_m$ | MAE   | $F_\beta$ | $S_m$ | MAE   | $F_\beta$ | $S_m$ | MAE   | $F_\beta$ | $S_m$ | MAE   | $F_\beta$ | $S_m$ | MAE   |
|--------------------|--------|--------|-----------|-------|-------|-----------|-------|-------|-----------|-------|-------|-----------|-------|-------|-----------|-------|-------|
| PSCNet (ours)      | ECSSD  | 2019   | 0.844     | 0.324 | 0.031 | 0.864     | 0.805 | 0.071 | 0.870     | 0.802 | 0.095 | 0.932     | 0.916 | 0.026 | 0.811     | 0.836 | 0.083 |
|                    | PASCAL-S | 2019   | 0.856     | 0.785 | 0.064 | 0.851     | 0.772 | 0.112 | 0.929     | 0.909 | 0.032 | 0.811     | 0.836 | 0.056 | 0.860     | 0.835 | 0.047 |
|                    | SOD    | 2019   | 0.870     | 0.802 | 0.095 | 0.932     | 0.916 | 0.026 | 0.875     | 0.870 | 0.037 | 0.875     | 0.870 | 0.037 | 0.875     | 0.870 | 0.037 |
|                    | HKU-IS | 2019   | 0.864     | 0.771 | 0.110 | 0.925     | 0.906 | 0.034 | 0.797     | 0.825 | 0.056 | 0.865     | 0.858 | 0.043 | 0.865     | 0.858 | 0.043 |
|                    | DUT-OMRON | 2019   | 0.844     | 0.771 | 0.110 | 0.925     | 0.906 | 0.034 | 0.797     | 0.825 | 0.056 | 0.865     | 0.858 | 0.043 | 0.865     | 0.858 | 0.043 |
|                    | DUTS-test | 2019   | 0.851     | 0.849 | 0.057 | -         | -     | -     | -         | -     | -     | -         | -     | -     | -         | -     | -     |

IV. EXPERIMENTAL RESULTS

This section describes six benchmark datasets and evaluation metrics used for salient object detection and compares our PSCNet with other 25 methods for salient object detection.

A. DATASETS AND EVALUATION METRICS

Six benchmark datasets for salient object detection are used in our experiments: (a) ECSSD [39] includes 1,000 images that have meaningful and complex structures. (b) PASCAL-S [59] includes 850 images that are selected from the PASCAL VOC2010 segmentation dataset [75] and there are several salient objects in each image. (c) SOD [60] includes 300 images that are selected from the BSDS dataset [76] and the salient objects have low contrast. (d) HKU-IS [61] includes 4,447 images and most of the images have multiple salient objects. (e) DUT-OMRON [40] includes 5,168 images with high quality and there are one or multiple salient objects in each image. (f) DUTS [62] includes a training set of 10,553 images and a testing set (denoted as DUTS-test) of 5,019 images, and we use the training set of DUTS as our training set to train the PSCNet by following the recent works on salient object detection [22], [24], [32], [36].

Then, three common metrics are used for the quantitative evaluation: F-measure ($F_\beta$), structure measure ($S_m$), and mean absolute error (MAE):

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

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

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

where $F_\beta$ is the combination of the precision and recall, and we follow [18], [77] to set $\beta^2 = 0.3$; $S_m$ [78] is used to compute the structural similarity between the predicted map $S$ and ground truth image $G$, and we use $S_o$ and $S_r$ to represent the object-aware and region-aware structural similarity and set $\alpha$ as 0.5; MAE denotes the average pixel-wise absolute difference between $S$ and $G$. In general, the larger the $F_\beta$ or $S_m$ and the smaller the MAE, the better the result. The implementation of $F_\beta$, $S_m$, and MAE in [18], [78] are used to evaluate all the results.

B. ABLATION STUDY

We perform the ablation study to evaluate the components in PSCNet. Here, we build two baseline networks: one is the Deeplab [56], which is constructed by removing the Pyramid Spatial Context Module from the overall network architecture; another is the ‘‘+SC’’, which is constructed by adopting a single two-round recurrent translation to aggregate the context information from the features extracted by feature extraction network. Table 1 shows the results, where after leveraging the two-round recurrent translation to aggregate the spatial context, we obtain an obvious improvement, and our PSCNet achieves the best performance on all the benchmark datasets by formulating the pyramid spatial con-
C. COMPARISON WITH THE STATE-OF-THE-ARTS

We compare our PSCNet with other 25 state-of-the-art methods for salient object detection; see the first column in Table 2 for the names of these methods. Among them, BSCA [69] and DRFI [9] use the hand-crafted features to detect salient objects and other methods all utilize deep convolutional neural networks to learn features for saliency detection. For a fair comparison, the results of the compared methods are obtained by using the saliency maps provided by authors or using their implementations to predict the saliency maps with the released training models.

1) QUANTITATIVE COMPARISON

Table 2 shows the quantitative comparison results with the 25 methods on salient object detection in terms of $F_\beta$, $S_m$, and MAE on six benchmark datasets. Our PSCNet achieves the best performance on all the benchmark datasets by comparing it with these 25 methods. Especially, our method shows a large improvement on the DUTS-test dataset, which includes many challenge cases, demonstrating the strong capability of our PSCNet. In the future, we will explore the potential of our PSC module design for other layer separation tasks, such as mirror detection [79], lane marking detection [80], shadow detection [81]–[83] and removal [84], [85], reflection removal [86], [87], rain removal [88], haze removal [89], [90], etc.

2) VISUAL COMPARISON

Figure 4 and Figure 5 present the visual comparison results on salient object detection. From these figures, we can see that other methods (d)-(h) may include the background noise or fail to find all of the salient objects. In contrast, our PSCNet is able to generate the saliency maps that are more consistent with the ground truth images (b). Particularly, our method can produce the accurate saliency maps in challenging cases, such as (i) small salient objects (see the second and fourth rows in Figure 4 and the third row in Figure 5), (ii) complex background (see the first, second, and last two rows in Figure 4 and the first, second, and fourth rows in Figure 5), and (iii) multiple objects (see the second to seventh rows in Figure 4 and the second and third rows in Figure 5).
V. CONCLUSION

This paper develops a novel deep network architecture, named as PSCNet, for salient object detection by exploring multiple levels of spatial context features. Our key idea is to aggregate multiple levels of spatial contexts on the feature maps with different respective fields to obtain the fine-grain and coarse-grain characteristics of the images. As a result, we can produce more discriminative features to reduce the non-salient noise, to enhance the salient details, and to improve saliency accuracy. In the end, we evaluate our PSCNet on six common benchmark datasets by comparing it with 25 methods for salient object detection. Experimental results show that our method outperforms all the others, both visually and quantitatively. In the future, we will explore the potential of our PSC module design for other computer vision tasks.

REFERENCES

[1] H. Hadizadeh and I. V. Bajic, “Saliency-aware video compression,” IEEE Trans. Image Process., vol. 23, no. 1, pp. 19–33, Jan. 2014.
[2] H. Zhao, X. Mao, X. Jin, J. Shen, F. Wei, and J. Feng, “Real-time saliency-aware video abstraction,” Vis. Comput., vol. 25, no. 11, pp. 973–984, Nov. 2009.
[3] M.-M. Cheng, F.-L. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, “RepFinder: Finding approximately repeated scene elements for image editing,” ACM Trans. Graph., vol. 29, no. 4, pp. 1–8, Jul. 2010.
[4] L. Zhu, X. Hu, C.-W. Fu, J. Qin, and P.-A. Heng, “Saliency-aware texture smoothing,” IEEE Trans. Vis. Comput. Graphics, early access, Dec. 21, 2018, doi: 10.1109/TVCG.2018.2889055.
[5] X. Hu, X. Xu, Y. Xiao, H. Chen, S. He, J. Qin, and P.-A. Heng, “SINet: A scale-insensitive convolutional neural network for fast vehicle detection,” IEEE Trans. Intell. Transp. Syst., vol. 20, no. 3, pp. 1010–1019, Mar. 2019.
[6] Y. Yu, X. Xu, X. Hu, and P.-A. Heng, “DALoNet: Improving localization accuracy for domain adaptive object detection,” IEEE Access, vol. 7, pp. 63155–63163, 2019.
[7] H. Zhang, J. Zhang, and P. Koniusz, “Few-shot learning via saliency-guided hallucination of samples,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 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 Trans. Pattern Anal. Mach. Intell., vol. 37, no. 3, pp. 569–582, Mar. 2015.
[9] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, “Saliency object detection: A discriminative regional feature integration approach,” in Proc. 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 Trans. Pattern Anal. Mach. Intell., vol. 33, no. 2, pp. 353–367, Feb. 2011.
[11] 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 Trans. Image Process., vol. 25, no. 8, pp. 3919–3930, Aug. 2016.
[12] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, “Non-local deep features for salient object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 6609–6617.
[13] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, “Saliency detection with recurrent fully convolutional networks,” in Proc. ECCV, 2016, pp. 825–841.
[41] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, “Learning uncertain convolutional features for accurate saliency detection,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 212–221.

[42] R. Zhao, W. Ouyang, H. Li, and X. Wang, “Saliency detection by multi-context deep learning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1265–1274.

[43] S. Chen, X. Tan, B. Wang, and X. Hu, “Reverse attention for salient object detection,” in Proc. ECCV, 2018, pp. 234–250.

[44] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, and P.-A. Heng, “R3Net: Recurrent residual refinement network for salience detection,” in Proc. 27th Int. Joint Conf. Artif. Intell. (IJCAI), Jul. 2018, pp. 684–690.

[45] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and H. P. Torr, “Deeply supervised salient object detection with short connections,” in Proc. CVPR, 2017, pp. 3203–3212.

[46] X. Hu, L. Zhu, J. Qin, C.-W. Fu, and P.-A. Heng, “Recursively aggregating deep features for salient object detection,” in Proc. AAAI, 2018, pp. 6943–6950.

[47] G. Li and Y. Yu, “Deep contrast learning for salient object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 478–487.

[48] X. Li, F. Yang, H. Cheng, W. Liu, and D. Shen, “Contour knowledge transfer for salient object detection,” in Proc. ECCV, 2018, pp. 370–385.

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

[50] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, “Anmulet: Aggregating multi-level convolutional features for salient object detection,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 202–211.

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

[52] P. Zhang, W. Liu, H. Lu, and C. Shen, “Salient object detection with lossless feature reflection and weighted structural loss,” IEEE Trans. Image Process., vol. 28, no. 6, pp. 3048–3060, Jun. 2019.

[53] W. Wang, S. Zhao, J. Shen, S. C. H. Hoi, and A. Borji, “Salient object detection with pyramid attention and salient edges,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 1448–1457.

[54] T. Zhao and X. Wu, “Pyramid feature attention network for saliency detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 3085–3094.

[55] Z. Wu, L. Su, and Q. Huang, “Cascade partial decoder for fast and accurate salient object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 3907–3916.

[56] 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 Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 3917–3926.

[57] 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 Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 5968–5977.

[58] 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 Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 8150–8159.

[59] 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/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 3127–3135.

[60] 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,” in Proc. IEEE Trans. Circuits Syst. Video Technol., early access, Sep. 13, 2019, doi: 10.1109/TCVSST.2019.2941017.

[61] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, “BASNet: Boundary-aware salient object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 7479–7489.

[62] N. Liu and J. Han, “A deep spatial contextual long-term recurrent convolutional network for saliency detection,” IEEE Trans. Image Process., vol. 27, no. 7, pp. 3264–3274, Jul. 2018.

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

[64] A. Borji and L. Itti, “Exploiting local and global patch rarities for saliency detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2012, pp. 478–485.

[65] V. Mahadevan and N. Vasconcelos, “Biologically inspired object tracking using center-surround saliency mechanisms,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 3, pp. 541–554, Mar. 2013.
[61] G. Li and Y. Yu, “Visual saliency based on multiscale deep features,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 5455–5463.

[62] 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 Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 136–145.

[63] L. Zhang, J. Zhang, Z. Lin, H. Lu, and Y. He, “CapSal: Leveraging captioning to boost semantics for salient object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 6024–6033.

[64] W. Wang, J. Shen, X. Dong, and A. Borji, “Salient object detection driven by fixation prediction,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 1711–1720.

[65] 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. (ICCV), Oct. 2017, pp. 4019–4028.

[66] N. Liu and J. Han, “DHSNet: Deep hierarchical saliency network for salient object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 678–686.

[67] G. Lee, Y.-W. Tai, and J. Kim, “Deep saliency with encoded low level distance map and high level features,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 660–668.

[68] 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. (CVPR), Jun. 2015, pp. 3183–3192.

[69] Y. Qin, H. Lu, Y. Xu, and H. Wang, “Saliency detection via cellular automata,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 110–119.

[70] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, “Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 2874–2883.

[71] X. Hu, C.-W. Fu, L. Zhu, J. Qin, and P.-A. Heng, “Direction-aware spatial context features for shadow detection and removal,” IEEE Trans. Pattern Anal. Mach. Intell., early access, May 2019, doi: 10.1109/TPAMI.2019.2919616.

[72] X. Hu, L. Zhu, C.-W. Fu, J. Qin, and P.-A. Heng, “Direction-aware spatial context features for shadow detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7454–7462.

[73] J. Hu, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImagNet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 248–255.

[74] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. ICLR, 2015, pp. 1–15.

[75] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The Pascal visual object classes (VOC) challenge,” Int. J. Comput. Vis., vol. 88, no. 2, pp. 303–338, Jun. 2010.

[76] 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. 8th IEEE Int. Conf. Comput. Vis. (ICCV), 2001, pp. 416–423.

[77] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, “Frequency-tuned salient region detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 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 Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 4548–4557.

[79] X. Yang, H. Mei, K. Xu, X. Wei, B. Yin, and R. Lau, “Where is my mirror?” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 8809–8818.

[80] X. Xu, T. Yu, X. Hu, W. W. Y. Ng, and P.-A. Heng, “SALMNet: A structure-aware lane marking detection network,” IEEE Trans. Intell. Transp. Syst., early access, Apr. 20, 2020, doi: 10.1109/TITS.2020.2983077.

[81] 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 Proc. ECCV, 2018, pp. 121–136.

[82] X. Hu, T. Wang, C.-W. Fu, Y. Jiang, Q. Wang, and P.-A. Heng, “Revisiting shadow detection: A new benchmark dataset for complex world,” 2019, arXiv:1911.06998. [Online]. Available: http://arxiv.org/abs/1911.06998

[83] T. Wang, X. Hu, Q. Wang, P.-A. Heng, and C.-W. Fu, “Instance shadow detection,” 2019, arXiv:1911.07034. [Online]. Available: http://arxiv.org/abs/1911.07034

[84] H. Le and D. Samaras, “Shadow removal via shadow image decomposition,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 8578–8587.

[85] X. Hu, Y. Jiang, C.-W. Fu, and P.-A. Heng, “Mask-ShadowGAN: Learning to remove shadows from unpaired data,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 2472–2481.

[86] R. Wan, B. Shi, L.-Y. Duan, A.-H. Tan, and A. C. Kot, “Benchmarking single-image reflection removal algorithms,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 3922–3930.

[87] Q. Wen, Y. Tan, J. Qin, W. Liu, G. Han, and S. He, “Single image reflection removal beyond linearity,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 3771–3779.

[88] X. Hu, C.-W. Fu, L. Zhu, and P.-A. Heng, “Depth-attentional features for single-image rain removal,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 8022–8031.

[89] Z. Deng, L. Zhu, X. Hu, C.-W. Fu, X. Xu, Q. Zhang, J. Qin, and P.-A. Heng, “Deep multi-model fusion for single-image dehazing,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 2453–2462.

[90] X. Liu, Y. Ma, Z. Shi, and J. Chen, “GridDehazeNet: Attention-based multi-scale network for image dehazing,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 7314–7323.

HUI LI received the M.S. degree from National Aviation University, Ukraine, in 2003. She is currently a Lecturer with the School of Computer Science and Engineering, Huizhou University. Her research interests include image processing, computer vision, and image recognition.

**H. Li:** PSC Features for Salient Object Detection.