Postprint: Salient Object Detection with Deep Reuse of Multi-Scale Features

Authors: Zhou Zhiping, Fan Bin, fir, Xu Wencheng

Date: 2022-04-07T15:01:57+00:00

Abstract

To address the limitations of conventional salient object detection methods in detecting multiple salient objects at different scales, we propose a salient object detection algorithm based on deep reuse of multi-scale features. The network model consists of vertically stacked bidirectional dense feature aggregation modules and horizontally stacked multi-resolution semantic complementation modules. First, the bidirectional dense feature aggregation module extracts semantic features at different resolutions based on the ResNet backbone network, then performs adaptive fusion sequentially along the top-down and bottom-up pathways to obtain multi-scale representation features at different levels; Finally, the multi-resolution semantic complementation module fuses multi-scale features from two adjacent levels to eliminate mutual interference between features at different levels and enhance the consistency of prediction results. Experimental results on five benchmark datasets demonstrate that the proposed method can achieve Fmax, Sm, and MAE values of up to 0.939, 0.921, and 0.028, respectively, with a detection speed of 74.6 FPS, exhibiting superior detection performance compared to other competing algorithms.

Full Text

Deep Multiplexing Multi-Scale Features for Salient Object Detection

Zhou Zhiping, Fan Bin, Gai Shan, Xu Wencheng
(School of Information Engineering, Nanchang Hangkong University, Nanchang 330063, China)

Abstract

Traditional salient object detection methods often struggle with detecting multiple salient targets at varying scales. To address this limitation, we propose

a novel salient object detection algorithm that deeply multiplexes multi-scale features. Our network architecture comprises vertically stacked bidirectional dense feature aggregation modules and horizontally stacked multi-resolution semantic complementary modules. First, the bidirectional dense feature aggregation module extracts semantic features at different resolutions based on the ResNet backbone, then adaptively fuses them along both top-down and bottom-up pathways to obtain multi-scale representation features at various hierarchical levels. Subsequently, the multi-resolution semantic complementary module fuses multi-scale features from two adjacent levels to eliminate mutual interference between features at different levels, thereby enhancing prediction consistency. Experimental results on five benchmark datasets demonstrate that our method achieves competitive performance with Fmax, Sm, and MAE scores reaching up to 0.939, 0.921, and 0.028 respectively, while maintaining a detection speed of 74.6 fps, outperforming existing state-of-the-art methods.

Keywords: salient object detection; multi-scale features; bidirectional dense feature aggregation; multi-resolution semantic; deep learning

0 Introduction

Salient object detection (SOD) is a critical technique in computer vision that aims to segment the most visually conspicuous regions from input images. The rise of deep learning has significantly advanced SOD technology, elevating its performance to new heights. SOD has been widely applied across various computer vision domains, including image segmentation [**?**], visual tracking [**?**], image quality assessment [**?**], image retrieval [**?**], and edge detection [**?**]. In CNN-based SOD models, features from different hierarchical levels characterize distinct properties of salient objects. Specifically, low-level semantic features contain detailed information about salient objects but suffer from substantial noise, while high-level semantic features help localize salient objects but lack detailed object information.

Two fundamental challenges remain unresolved: how to extract more effective information from scale-varying data, and how to maintain spatial consistency between predictions and salient targets in images. Recent research has focused on designing sophisticated network architectures to extract highly discriminative multi-scale features or to fuse them efficiently, thereby meeting the demands of detecting salient objects at different scales. Zhang et al. [**?**] proposed a general framework for aggregating multi-level convolutional features, combining features from multiple layers in a fully connected manner. Hou et al. [**?**] introduced skip connections into the Holistically-Nested Edge Detection (HED) model, proposing a skip-layer architecture with a series of shortcut connections from high-level to low-level features. Liu et al. [**?**] constructed global context features by selectively aggregating contextual information, then merged global and multi-scale local contexts to improve performance. Wu et al. [**?**] proposed a novel cascaded partial decoder framework that discards low-level features to reduce complexity in deep aggregation models and utilizes generated relatively

accurate attention maps to refine high-level features. Pang et al. [?] introduced an aggregation interaction module that effectively leverages adjacent layer features through mutual learning and adaptive modules, enabling the network to adaptively extract multi-scale information for better handling scale variations.

Furthermore, merging multi-level features is essential for generating better saliency maps. However, excessive integration of features at different resolutions not only incurs substantial computational overhead but also dilutes useful features, leading to degraded algorithm performance. To overcome this issue, researchers have proposed various solutions. Feng et al. [?] employed attention feedback modules constructed from each encoder block and its corresponding decoder block to help combine multi-level features. Wei et al. [?] adopted a selective fusion strategy that merges features from different levels through element-wise multiplication to suppress redundant information and avoid cross-contamination between hierarchical features. Qin et al. [?] proposed a two-level nested U-structure to integrate deep features from multiple levels. Chen et al. [?] introduced residual learning into the HED architecture, using reverse attention in the top-down pathway to guide residual saliency learning, enabling the network to quickly and effectively discover missing object parts and details. Chen et al. [?] combined a center-surround contrast mechanism with convolutional neural networks, providing a powerful approach to effectively enhance the representation capability of multi-scale features.

In summary, effectively fusing features from different hierarchical levels in CNN backbone networks is crucial. Building upon the U-Net model [?], this paper proposes a salient object detection model called DMMF (Deep Multiplexing Multi-scale Feature). The model introduces a Bidirectional Dense Aggregation module (BDA) that reuses CNN features extracted from the backbone network at different resolutions along both top-down and bottom-up pathways, employing residual connections for feature enhancement. Multiple BDA modules at different scales are stacked to extract multi-level features with multi-resolution semantics. Drawing inspiration from [?], we design a Multi-resolution Semantic Complement module (MSC) that is cascaded into the bottom-up pathway of the U-Net architecture to enhance the model's predictive capability for salient objects.

1 Deep Multiplexing Multi-Scale Feature Network

The overall architecture of our deep multiplexing multi-scale feature network for salient object detection is illustrated in Figure 1. Using ResNet-50 as the backbone network, we propose a stacked bidirectional dense feature aggregation module to perform full-resolution fusion of the extracted features, yielding more semantically rich multi-scale features. For the obtained multi-level multi-scale features, cascaded multi-resolution semantic complementary modules are employed to preserve useful information from adjacent feature nodes, progressively restoring the semantic and spatial information of salient objects.

1.1 Bidirectional Dense Feature Aggregation Module

The bidirectional dense feature aggregation module aims to aggregate features at different resolutions. Formally, given a multi-scale feature list $\langle\langle MATH_1\rangle\rangle$, where $\langle\langle MATH_2\rangle\rangle$ represents features at level $\langle\langle MATH_3\rangle\rangle$, the algorithm seeks to find a transformation $\langle\langle MATH_4\rangle\rangle$ that can effectively aggregate different features and output a new feature list. Traditional FPN [**?**] aggregates multi-scale features in a top-down manner, which is essentially limited by unidirectional information flow. To address this limitation, PANet [**?**] adds an additional bottom-up path aggregation network, improving performance but introducing more parameters and computation. NAS-FPN [**?**] uses neural architecture search to find better cross-scale feature network topologies, but requires thousands of GPU hours during search and produces irregular, difficult-to-interpret or modify networks. EfficientDet [**?**] simplifies PANet to construct the BiFPN module and stacks it multiple times to more effectively obtain more discriminative multi-scale features.

The bidirectional dense feature aggregation module extracts multi-scale features through bidirectional (top-down and bottom-up) cross-scale connection paths. When fusing features at different resolutions, since input features have varying resolutions, they typically contribute unequally to the output features. The algorithm adds an extra weight to each input through a simple attention mechanism, allowing the network to learn the importance of each input feature. However, unlike simple concatenation operations, this algorithm achieves higher-level feature fusion by stacking bidirectional dense feature aggregation modules with progressively decreasing scales, achieving the same or even better effects with fewer parameters.

The structure of the BDA module is shown in Figure 2. Taking BDA5 in Figure 2 as an example, the subscript "5" indicates that this module has 5 input signals, corresponding to the basic features $\langle\langle MATH_5\rangle\rangle$ extracted from the 5 stages of the ResNet-50 network. First, $\langle\langle MATH_6\rangle\rangle$ is obtained through element-wise linear weighted operations and 3$×$3 convolution (including batch normalization and ReLU activation) on $\langle\langle MATH_7\rangle\rangle$ after upsampling. Similarly, $\langle\langle MATH_8\rangle\rangle$, $\langle\langle MATH_9\rangle\rangle$, and $\langle\langle MATH_{10}\rangle\rangle$ are obtained sequentially from bottom to top. Then, $\langle\langle MATH_{11}\rangle\rangle$ is obtained through element-wise linear weighted operations and 3$×$3 convolution on $\langle\langle MATH_{12}\rangle\rangle$ after downsampling and $\langle\langle MATH_{13}\rangle\rangle$. Similarly, $\langle\langle MATH_{14}\rangle\rangle$, $\langle\langle MATH_{15}\rangle\rangle$, and $\langle\langle MATH_{16}\rangle\rangle$ are obtained sequentially from top to bottom. Finally, $\langle\langle MATH_{17}\rangle\rangle$ serves as one input to MSC, while $\langle\langle MATH_{18}\rangle\rangle$, $\langle\langle MATH_{19}\rangle\rangle$, $\langle\langle MATH_{20}\rangle\rangle$, and $\langle\langle MATH_{21}\rangle\rangle$ serve as inputs to module BDA4. Similar to BDA5, BDA4 produces four outputs $\langle\langle MATH_{22}\rangle\rangle$, $\langle\langle MATH_{23}\rangle\rangle$, $\langle\langle MATH_{24}\rangle\rangle$, and $\langle\langle MATH_{25}\rangle\rangle$, where $\langle\langle MATH_{26}\rangle\rangle$ serves as one input to MSC, while $\langle\langle MATH_{27}\rangle\rangle$, $\langle\langle MATH_{28}\rangle\rangle$, and $\langle\langle MATH_{29}\rangle\rangle$ serve as inputs to module BDA3. Ultimately, the three outputs of BDA3 become inputs $\langle\langle MATH_{30}\rangle\rangle$, $\langle\langle MATH_{31}\rangle\rangle$, and $\langle\langle MATH_{32}\rangle\rangle$ to MSC.

The fusion process of BDA5 in Figure 2 is described by Equations (1) and (2):

$$\langle\langle MATH_{33}\rangle\rangle$$

where $\langle\langle MATH_{34}\rangle\rangle$, $\langle\langle MATH_{35}\rangle\rangle$ are intermediate features in the top-down pathway, *conv* denotes a set of operations including 3$×$3 convolution, batch normalization, and ReLU activation, $\langle\langle MATH_{36}\rangle\rangle$ are weight coefficients assigned to each input during feature fusion, initialized as random numbers in $(0, 1)$ and normalized using Laplace smoothing:

$$\langle\langle MATH_{37}\rangle\rangle$$

where $\langle\langle MATH_{38}\rangle\rangle$ is used to avoid numerical instability. The network updates $\langle\langle MATH_{39}\rangle\rangle$ after each training iteration, uses the ReLU function to ensure non-negativity, and re-normalizes through Laplace smoothing.

1.2 Multi-Resolution Semantic Complement Module

While the stacked bidirectional dense feature aggregation module extracts effective multi-scale features at different levels from the backbone network, the multi-resolution semantic complement module enables adjacent-level multi-scale features to complement each other spatially and semantically. This process continuously enhances features suitable for the current resolution while weakening unsuitable ones, thereby finding features appropriate for the current input information.

Through BDA5 to BDA3 in the top-down pathway, we obtain a final set of multi-scale semantic features $\langle\langle MATH_{40}\rangle\rangle$ with resolutions of $320\$\times 320, 160\times 160, 80\times 80, 40\times 40, and 20\times\$20$, respectively. These features all contain semantic information from targets of different scales, but the importance of each semantic component varies. Simply linearly fusing these features would cause mutual interference, weakening features with strong discriminative power and consequently degrading detection performance. To address this, we propose the multi-resolution semantic complement module to fully mine useful information from features at each granularity, creating complementary advantages. The detailed structure of MSC is shown in Figure 3.

MSC can be expressed as:

$$\langle\langle MATH_{41}\rangle\rangle$$

where $\langle\langle MATH_{42}\rangle\rangle$ and $\langle\langle MATH_{43}\rangle\rangle$ represent adjacent features, *conv* denotes dilated convolution with batch normalization and ReLU activation,

$\langle\langle MATH_{44}\rangle\rangle$ is element-wise addition, and $\langle\langle MATH_{45}\rangle\rangle$ is element-wise multiplication. MSC first merges input features through concatenation, then applies dilated convolution with rate 1, followed by normalization and ReLU operations to obtain global semantic information fused from both input features. This global semantic information is then added back to the input features through element-wise addition to enhance them spatially and semantically. Finally, element-wise multiplication with adaptive weights selectively inherits the two groups of enhanced features. Thus, MSC enables input features to inherit important characteristics while discarding more noise.

2 Experiments

2.1 Datasets

Training Dataset: The proposed method is trained on DUTS-TR, a subset of the DUTS dataset containing 10,553 images. It is currently the largest and most commonly used training dataset for salient object detection. To ensure model convergence, we set the training epochs to 80, using SGD optimizer with an initial learning rate of 1×10^{-3}, weight decay of 5×10^{-4}, and momentum coefficient of 0.9. All experiments are conducted on Linux 16.04 OS with GTX TITAN-XP GPU, PyTorch 1.0.0, and CUDA 9.0.

Testing Datasets: We evaluate the proposed method on six widely-used benchmark datasets: DUT-OMRON [**?**], DUTS-TE [**?**], HKU-IS [**?**], ECSSD [**?**], and PASCAL-S [**?**]. DUT-OMRON contains 5,168 images, most featuring one or two foreground objects with complex structures. The DUTS dataset comprises DUTS-TR and DUTS-TE. Since DUTS-TR is used for training, we select DUTS-TE with 5,019 images for testing. HKU-IS contains 4,447 images with multiple discontinuous salient objects intersecting image boundaries. ECSSD includes 1,000 images with complex structures, most containing large-scale foreground objects. PASCAL-S comprises 850 images with complex foreground objects and cluttered backgrounds.

2.2 Evaluation Metrics

For comprehensive evaluation, we employ three widely-used metrics: F-measure, Mean Absolute Error (MAE), and S-measure.

F-measure is the weighted harmonic mean of Precision and Recall, defined as:

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

where β^2 is generally set to 0.3. Higher F-measure indicates more accurate predictions. We report the maximum value across all thresholds as the evaluation result.

Mean Absolute Error (MAE) calculates the average absolute difference between predicted saliency maps and ground truth:

$$\mathrm{MAE} = \frac{1}{H \times W} \sum_{x=1}^{H} \sum_{y=1}^{W} |P(x,y) - G(x,y)|$$

where P represents the predicted saliency map, G the ground truth, (H, W) the image dimensions, and (x, y) pixel coordinates. Lower MAE indicates better performance.

S-measure evaluates structural similarity between predicted and ground truth saliency maps by measuring object-aware and region-aware structural similarity:

$$S_\alpha = \alpha \times S_o + (1 - \alpha) \times S_r$$

where α is typically set to 0.5. Higher S-measure indicates greater spatial structural similarity between detected saliency maps and ground truth.

2.3 Performance Analysis

We compare our proposed method with 11 state-of-the-art methods on five salient object detection datasets (DUTS-TE, DUT-OMRON, HKU-IS, ECSSD, and PASCAL-S) using the aforementioned metrics, as shown in Table 1.

Table 1. Comparison of $F_{\max}$, S_m, and MAE across different algorithms

Approach	DUTS-TE	DUT-OMRON	HKU-IS	ECSSD	PASCAL_S
	$F_{\max} \uparrow$ $S_m \uparrow$ MAE$\downarrow$	$F_{\max} \uparrow$ $S_m \uparrow$ MAE$\downarrow$	$F_{\max} \uparrow$ $S_m \uparrow$ MAE$\downarrow$	$F_{\max} \uparrow$ $S_m \uparrow$ MAE$\downarrow$	$F_{\max} \uparrow$ $S_m \uparrow$ MAE$\downarrow$
MWS[26]					
RAS[14]					
R3Net[27]					
CPD[9]					
AFNet[11]					
PoolNet[28]					
BASNet[30]					
EGNet[29]					
U2Net[13]					
F3Net[12]					
MINet[10]					

As shown in Table 1, our method demonstrates excellent performance, surpassing other salient object detection models on most datasets and metrics. Specifi-

cally, on HKU-IS, our algorithm achieves the best results across all three metrics: $F_{\max}$ improves by 0.004, S_m by 0.005, and MAE decreases by 0.003 compared to U2Net. On DUTS-TE, our method outperforms others in $F_{\max}$ and S_m, with only MAE slightly lower than F3Net. On ECSSD, our method achieves the best MAE and S_m, with only $F_{\max}$ slightly lower than U2Net. Overall, across all datasets and metrics, our method exhibits strong performance for multiple salient objects and scale variations.

Additionally, Table 2 compares the average speed (FPS) of different methods on the ECSSD dataset.

Table 2. Comparison of detection speed (FPS) across different algorithms

Algorithm	Input Size	FPS
R3Net	320$×320|||AFNet|256×256|||PoolNet|224×224|||BASNet|352×352|||EGNet	74.6

Our algorithm achieves 74.6 FPS, second only to MINet's 86 FPS, while delivering superior detection performance.

For more intuitive demonstration of our algorithm's advantages, Figure 4 visualizes prediction results from 11 state-of-the-art methods across different scenarios. The first row shows small salient targets, the second row large targets, rows 3-4 multiple targets of different sizes, row 5 low foreground-background contrast, and row 6 complex scenes.

As shown in Figure 4, our algorithm's detection results closely match ground truth across small targets, large targets, multi-scale targets, and complex backgrounds. Compared to other methods, our approach effectively suppresses background interference to detect small salient targets, produces more complete detection of large salient objects without missing parts, accurately captures object contours for multi-scale targets, successfully identifies complete object contours under low contrast conditions, and effectively detects salient targets without interference in complex scenes.

2.4 Ablation Analysis

a) Impact of Different BDA Module Stacking Strategies

To obtain better multi-scale information, our algorithm strategically stacks multiple BDA modules. To determine the optimal stacking configuration, we test and compare different schemes on the HKU-IS dataset, as shown in Table 3, where "$\times n$" denotes stacking the same structure n times.

Table 3. Performance comparison of different stacking methods

Stacking Strategy	$F_{\max}$	S_m	MAE
BDA5$\times$2\|\|\|\|\|$BDA5\times3$\|\|\|\|\|$BDA5\times$\$4			
BDA5+BDA4+BDA3			**Best**

The results show that stacking BDA5 multiple times outperforms using a single module, indicating that repeated fusion of multi-resolution features improves performance. However, performance degrades when stacking exceeds 3 times, as deeper networks suffer from gradient vanishing, making training more difficult. Stacking three BDA modules with different configurations achieves optimal performance, demonstrating that fusing multi-level semantic features at different resolutions prevents useful features from being diluted and captures more discriminative abstract features.

b) Impact of MSC and BDA Combination Strategies

To validate the effectiveness of MSC and BDA modules, we test different combination strategies on DUTS-TE using F-measure, MAE, and S-measure for comparison, as shown in Table 4.

Table 4. Ablation experiment of the proposed algorithm

Configuration	$F_{\max}$	S_m	MAE
Baseline			
Baseline + BDAs*			
Baseline + MSC			
Baseline + BDAs + MSC	**Best**	**Best**	**Best**

Here, **Baseline** denotes the original U-Net model, **BDAs** represents the subnetwork stacked as "BDA5+BDA4+BDA3", and **BDAs*** denotes unweighted BDAs where all weights $\langle\langle MATH_{46}\rangle\rangle$ in Equations (1) and (2) are set to 1. The results show that introducing either BDAs or MSC to the baseline improves performance, while the **Baseline+BDAs+MSC** strategy achieves the best results, with $F_{\max}$ and S_m improving by 0.038 and 0.048 respectively, and MAE decreasing by 0.02 compared to the baseline. This demonstrates that stacking multiple BDA and MSC modules extracts more abstract features beneficial for detection, and introducing adaptive weighting strategies to fuse hierarchical features prevents mutual interference, making predictions more consistent with salient objects in images.

3 Conclusion

To address multi-scale challenges in salient object detection, we propose a method based on deep multiplexing of multi-scale features. The method designs a bidirectional dense feature aggregation module that repeatedly

reuses convolutional features from the backbone network, employing adaptive weighted fusion to eliminate mutual interference between hierarchical features. Additionally, a multi-resolution semantic complement module fuses two sets of adjacent-resolution features for mutual spatial and semantic enhancement. Experimental results demonstrate that our method achieves $F_{\max}$, S_m, and MAE scores of 0.939, 0.921, and 0.028 respectively, outperforming 11 state-of-the-art methods. Our approach accurately detects multiple targets at different scales and effectively handles complex background scenarios. Future work will incorporate multi-supervision ideas and novel attention mechanisms to more robustly locate salient object contours, and employ depthwise separable convolutions to reduce model parameters.

References

[1] Li Fenglin, Li Liang. Target Image Segmentation Algorithm Based on Saliency Detection [J]. Electronic Science and Technology, 2017, 30(1): 69-71.

[2] Wang Yong, Wei Xian, Lu Ding, et al. A robust visual tracking method via local feature extraction and saliency detection [J]. The Visual Computer, 2020, 36(4): 683-700.

[3] Chen Chen. Research on Visual Perception-Based Image Quality Assessment Methods [D]. Xi'an: Xidian University, 2019.

[4] Wang Haoxiang, Li Zhihui, Yang Li, et al. Visual saliency guided complex image retrieval [J]. Pattern Recognition Letters, 2020, 130(2020): 64-72.

[5] Zhang Yanbang, Zhang Fen, Zhang Jiaojiao. Target detection algorithm based on image edge features [J]. Neijiang Science and Technology, 2021, 42(04): 47-67.

[6] Zhang Pingping, Wang Dong, Lu Huchuan, et al. Amulet: Aggregating multi-level convolutional features for salient object detection [C]// Proc of the IEEE International Conference on Computer Vision. IEEE: MIT Press, 2017: 202-211.

[7] Hou Qibin, Cheng Mingming, Hu Xiaowei, et al. Deeply supervised salient object detection with short connections [C]// Proc of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE: MIT Press, 2017: 3203-3212.

[8] Liu Nian, Han Junwei, Yang Ming-Hsuan. Picanet: Learning pixel-wise contextual attention for saliency detection [C]// Proc of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE: MIT Press, 2018: 3089-3098.

[9] Wu Zhe, Li Su, Huang Qingming. Cascaded partial decoder for fast and accurate salient object detection [C]// Proc of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE: MIT Press, 2019: 3907-3916.

[10] Pang Youwei, Zhao Xiaoqi, Zhang Lihe, et al. Multi-scale interactive network for salient object detection [C]// Proc of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE: MIT Press, 2020: 9413-9422.

[11] Feng Mengyang, Lu Huchuan, Ding Errui. Attentive feedback network for boundary-aware salient object detection [C]// Proc of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE: MIT Press, 2019: 1623-1632.

[12] Wei Jun, Wang Shuhui, Huang Qingming. F^3Net: Fusion, Feedback and Focus for Salient Object Detection [C]// Proc of the AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2020: 12321-12328.

[13] Qin Xuebin, Zhang Zichen, Huang Chenyang, et al. U2-Net: Going deeper with nested U-structure for salient object detection [J]. Pattern Recognition, 2020, 106: 107404.

[14] Chen Shuhan, Tan X, Wang Ben, et al. Reverse attention for salient object detection [C]// Proc of the European Conference on Computer Vision (ECCV). Berlin: Springer, 2018: 234-250.

[15] Chen Qin, Zhu Lei, Hou Yunlong, et al. Salient object detection based on depth center neighborhood pyramid structure [J]. Pattern Recognition and Artificial Intelligence, 2020, 33(06): 496-506.

[16] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation [C]// International Conference on Medical image computing and computer-assisted intervention. Berlin: Springer, 2015: 234-241.

[17] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection [C]// Proc of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE: MIT Press, 2017: 2117-2125.

[18] Mei Yiqun, Fan Yuchen, Zhang Yulun, et al. Pyramid attention networks for image restoration [J]. arXiv preprint arXiv: 2004.13824, 2020.

[19] Ghiasi G, Lin T Y, Le Q V. Nas-fpn: Learning scalable feature pyramid architecture for object detection [C]// Proc of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE: MIT Press, 2019: 7036-7045.

[20] Tan Mingxing, Pang Ruoming, Le Q V. Efficientdet: Scalable and efficient object detection [C]// Proc of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE: MIT Press, 2020: 10781-10790.

[21] Yang Chuan, Zhang Lihe, Lu Huchuan, et al. Saliency detection via graph-based manifold ranking [C]// Proc of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE: MIT Press, 2013: 3166-3173.

[22] Wang Lijun, Lu Huchuan, Wang Yifan, et al. Learning to detect salient objects with image-level supervision [C]// Proc of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE: MIT Press, 2017: 3794-3802.

[23] Li Guanbin, Yu Yizhou. Visual saliency based on multiscale deep features [C]// Proc of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE: MIT Press, 2015: 5455-5463.

[24] Yan Qiong, Xu Li, Shi Jianping, et al. Hierarchical saliency detection [C]// Proc of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE: MIT Press, 2013: 1155-1162.

[25] Li Yin, Hou Xiaodi, Koch C, et al. The secrets of salient object segmentation [C]// Proc of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE: MIT Press, 2014: 280-287.

[26] Zeng Yu, Zhuge Y, Lu Huchuan, et al. Multi-source weak supervision for saliency detection [C]// Proc of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE: MIT Press, 2019: 6074-6083.

[27] Deng Zijun, Hu Xiaowei, Zhu Lei, et al. R3net: Recurrent residual refinement network for saliency detection [C]// Proc of the 27th International Joint Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2018: 684-690.

[28] Liu Jiangjiang, Hou Qibin, Cheng Mingming, et al. A simple pooling-based design for real-time salient object detection [C]// Proc of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE: MIT Press, 2019: 3917-3926.

[29] Zhao Jiaxing, Liu Jiangjiang, Fan Dengping, et al. EGNet: Edge guidance network for salient object detection [C]// Proc of the IEEE/CVF International Conference on Computer Vision. 2019: 8779-8788.

[30] Qin Xuebin, Zhang Zichen, Huang Chenyang, et al. Basnet: Boundary-aware salient object detection [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE: MIT Press, 2019: 7479-7489.

Note: Figure translations are in progress. See original paper for figures.