Hindawi
Advances in Multimedia
Volume 2022, Article ID 2243927, 14 pages
https://doi.org/10.1155/2022/2243927

Research Article

Multiscale Deep Network with Centerness-Aware Loss for Salient Object Detection

Liangliang Duan

Qingdao University of Technology, Qingdao, Shandong 26600, China

Correspondence should be addressed to Liangliang Duan; hengxingdll9@163.com

Received 10 July 2021; Revised 29 August 2021; Accepted 23 November 2021; Published 13 January 2022

Academic Editor: Patrick Seeling

Deep encoder-decoder networks have been adopted for saliency detection and achieved state-of-the-art performance. However, most existing saliency models usually fail to detect very small salient objects. In this paper, we propose a multitask architecture, M2Net, and a novel centerness-aware loss for salient object detection. The proposed M2Net aims to solve saliency prediction and centerness prediction simultaneously. Specifically, the network architecture is composed of a bottom-up encoder module, top-down decoder module, and centerness prediction module. In addition, different from binary cross entropy, the proposed centerness-aware loss can guide the proposed M2Net to uniformly highlight the entire salient regions with well-defined object boundaries. Experimental results on five benchmark saliency datasets demonstrate that M2Net outperforms state-of-the-art methods on different evaluation metrics.

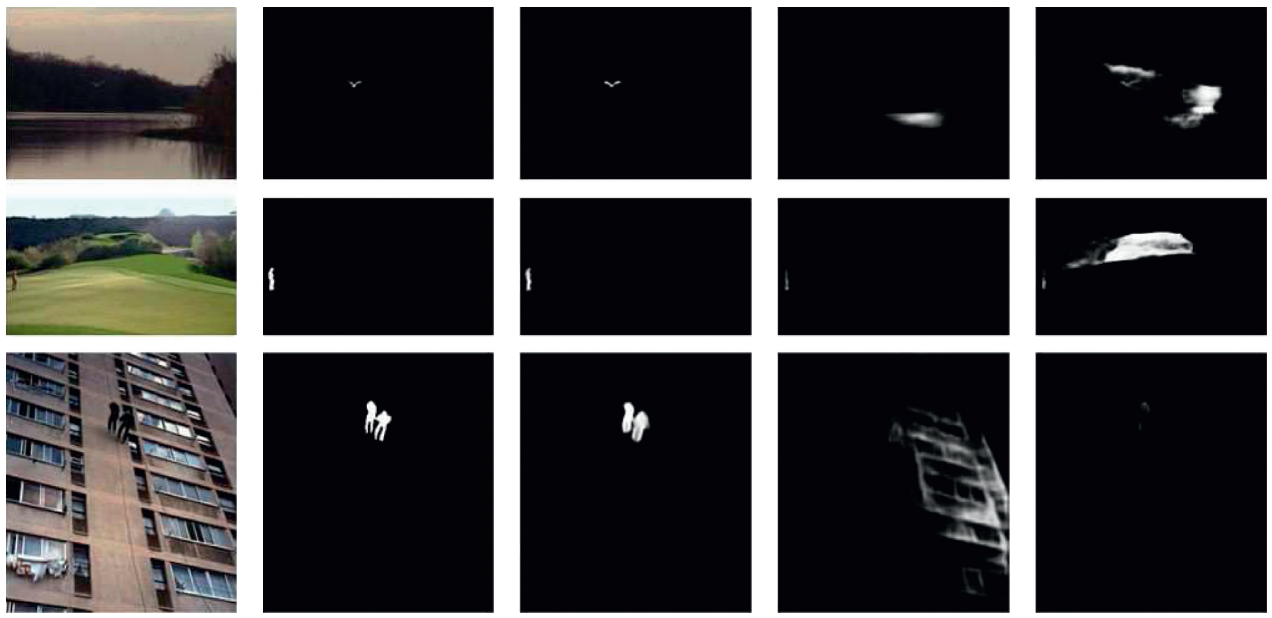
1. Introduction

Salient object detection (SOD) [1–3] aims to extract the most visually distinctive objects in an image or video. During the past decades, it has become a hotspot in the research field of computer vision. Saliency detection results often serve as the first step for a variety of downstream computer vision tasks, including object recognition [4], visual tracking [5], image retrieval [6], no-reference synthetic image quality assessment [7], robot navigation [8] image and video compression [9, 10], and object discovery [11–13].

Earlier SOD methods mostly rely on hand-crafted features (e.g., color, brightness, and texture) to produce saliency maps. However, these low-level features can hardly capture high-level semantic information and are not robust enough to various complex scenarios.

Recently, convolutional neural networks (CNNs), especially fully convolutional neural networks (FCNs) [14], have pushed salient object detection to achieve very promising results on many popular public benchmark datasets. Encoder-decoder framework [3, 15–19] is frequently used to extract and combine enriched feature blocks and therefore can generate more accurate saliency maps. More recently, many researchers further improved the saliency model by incorporating domain-specific information from other tasks such as contour/edge detection [18, 20, 21], image classification [22, 23], and noise pattern modeling [24].

These U-shape models [3, 21] have greatly refreshed the leaderboards on all commonly used datasets. However, existing saliency methods still hold many problems that are not solved totally and are worthy of further research. First, due to the repeated subsampling, a single-scale convolutional kernel has difficulty in accurately segmenting size-varying salient object. Two state-of-the-art methods cannot uniformly highlight small foreground object with well-defined boundaries, as is shown in Figures 1(d) and 1(e). This motivates some efforts to characterize the multiscale information from a single layer. Second, most of the existing saliency methods [15, 25] use binary cross entropy (BCE) loss to train the saliency networks. But these models with BCE loss usually have low confidence in making a distinction between foreground and background, leading to blurred boundaries. The recent survey [26] indicates that the elaborate design of loss function can help to train more effective saliency detection models. Some training losses, such as PPA



(a) (b) (c) (d) (e)

Figure 1: Sample results of our approach (M2Net) compared to GateNet and MINet. (a) and (b) show the input images with small foreground object and the ground truth (GT), respectively. (c), (d), and (e) are saliency maps of ours (M2Net), GateNet, and MINet.

loss [19], Intersection over Union (IoU) loss [17, 27], and F-measure loss [28], were proposed for improving model performance. In consequence, it is essential to design a mechanism to extract multiscale information from each layer and develop a novel training loss.

To address the above challenges, we proposed a novel multiscale and multitask network, named M2Net, which can generate high-quality saliency maps with clear boundaries (see Figure 1(c)). Firstly, in the bottom-up encoder module, we use two branches to extract robust feature blocks. The backbone branch is based on a common pretrained image classification network, while the transformation branch is based on the sequence of three operations, including convolution, batch normalization, and ReLU. Secondly, in the decoder module, we develop two units, including multiscale feature extraction unit and cross-layer feature block fusion unit, to generate the saliency maps. Multiscale feature extraction unit can extract multiscale contextual features, while cross-layer feature block fusion unit can continually fuse adjacent level feature blocks. Thirdly, to take full advantage of ground truth, we design a centerness-aware loss, which considers the location of salient objects. This loss can guide the proposed network to generate high-quality saliency maps.

We conduct experiments on five benchmark saliency datasets and demonstrate the better performance of the proposed M2Net. In summary, our contributions are as follows:

(i) We propose a multiscale and multitask deep framework with a centerness-aware loss for salient object detection. The M2Net consists of encode module, decoder module, and centerness prediction module.

(ii) We develop a centerness-aware loss, which can help to generate high-quality saliency maps, and it can push the proposed M2Net to uniformly highlight the entire salient regions with clear boundaries.

(iii) Extensive experiments on five public SOD datasets show that our model M2Net outperforms state-of-the-art saliency methods on different evaluation metrics. In particular, the proposed model (M2Net) can achieve the best performance under different challenging situations.

2. Related Work

2.1. Salient Object Detection. Early SOD methods [2, 29, 30] are mainly based on hand-crafted features and some intrinsic cues, such as center prior, color contrast, and background prior. Recently, convolutional neural networks (CNNs) have been used to extract multilevel features from input images. CNNs-based methods treat patches/superpixels [31–33] and generic object proposals [34–37] as image processing units, and an MLP-classifier is used to train the network. Wang et al. [35] trained two different CNN models to generate a saliency map. DNN-L and DNN-G are used to extract local and global features, respectively. Particularly, fully convolutional networks (FCN) show their advantage and refresh the state-of-the-art records in saliency prediction task. The encoder-decoder framework is frequently used in the FCN-based saliency models [3, 15–19, 38–40]. Liu et al. [16] proposed a novel network to embed local and global pixelwise contextual attention modules into a U-shape network. Zhao et al. [3] proposed a simple and effective gated network architecture to control the meaningful message

passing from encoder to decoder feature blocks. Almost all of the above methods try to develop more complicated modules and strategies to fuse feature blocks of different levels. Different from the methods mentioned above, we propose a simple and effective multitask architecture, which attempts to solve saliency tasks by adding an extra centerness prediction branch.

2.2. Multiscale Feature Extraction. The atrous spatial pyramid pooling (ASPP) module [41] is widely used in many computer vision tasks. The atrous convolution can expand the receptive field with fewer parameters to get large-scale and more comprehensive features. The pyramid pooling module (PPM) [42] is another choice for extracting multiscale features. Zhang et al. [43] insert five ASPP modules into the encoder feature blocks of five levels. The larger the atrous rate, the more the difficulty in capturing the changes of image details. To alleviate the above problem, Zhao et al. [3] designed a folded ASPP and achieved a local-in-local effect. Besides, the pyramid attention module [44] can generate multiscale attention maps to enhance saliency features. The above methods can extract multiscale features from images, but it is more sensitive to background noise. To improve the recall rate of saliency objects under complex background, we propose a multiscale feature extraction module and insert it into decoder feature blocks.

2.3. Multitask Learning. Multitask learning (MTL) has led to successes in many research fields, from computer vision and speech recognition to drug discovery and natural language processing. Multitask learning aims at simultaneous training using two or more related tasks. It is found that learning multiple tasks jointly can lead to better performance improvement compared with learning them individually. Recent multitask learning-based saliency methods have shown good results by jointly tackling multiple related tasks such as image classification, fixation prediction, and edge detection. Li et al. [23] and Wang et al. [22] proposed to apply image-level tags to assist the detection of the foreground object. Kruthiventi et al. [47] proposed a unified multitask learning framework to jointly solve salient object detection and fixation prediction. Zhao et al. [20] presented an edge guidance network to extract two complementary features, including salient object features and salient edge features. As we all know, location is the important information of an object. To the best of our knowledge, this information has never been directly used in saliency prediction tasks. In this paper, we investigate how to integrate the centerness prediction task into saliency detection.

3. Proposed Method

In this paper, we propose a multitask and multiscale deep network for salient object detection. The overview of the proposed network consists of three related modules, as shown in Figure 2. To guide the saliency network to uniformly highlight the entire object with different size, we propose a multiscale feature extraction approach. To further improve the detection accuracy, we introduce centerness-aware loss, which helps to reduce the impact of complex background.

3.1. Network Overview. The encoder-decoder architecture has been widely used in the salient object detection task, and it has a strong ability to combine features from different network layers. Our method is built on the feature pyramid networks (FPN) [48] with the pretrained ResNeXt-101 [46] or ResNet-50 [45] as the backbone network, both of which can extract meaningful saliency features to build high-quality U-shape networks. To reduce network parameters, we discard all the fully connected layer of the pretrained backbone [45, 46]. The proposed M2Net consists of a bottom-up encoder module, top-down decoder module, and centerness prediction module. In the encoder, we use the pretrained backbone to extract multilevel saliency features from preprocessed images. To obtain robust saliency features, each feature block is processed by 1×1 convolutional layers followed by batch norm and ReLU (Figure 2). Next, in the decoder, we use a skip/concatenation connection scheme. To generate the final saliency maps, a novel multiscale feature extraction approach is proposed (Figure 3). Lastly, we design a centerness prediction module (Figure 2), which can help to generate high-quality saliency maps. We describe the structures of the three modules and explain their transformation in the following sections.

3.2. Encoder Module. In our M2Net, the encoder module is composed of a backbone branch and a transformation branch. The backbone branch is based on a common pretrained image classification network, for example, the VGG, ResNet-50 [45] or ResNeXt-101 [46]. In order to fit the saliency prediction task, similar to most previous saliency methods [3, 16, 20], we remove the last pooling layer and cast away all the fully connected layers of the ResNet or ResNeXt. Let $I \in R^{320 \times 320 \times C}$ denote the input training image with ground truth labels $Y \in R^{320 \times 320 \times 1}$ as is shown in Figure 2, where C denotes the channel of the input image. For a given input image with size $H \times W$, the pretrained image classification network will extract its saliency features at five different levels, denoted as $\{E^i \in R^{H \times W \times C} | i = 1, \ldots, 5\}$ with resolutions $[H/(2^{i-1}), W/(2^{i-1})]$, where C denotes the channel of the feature blocks. H and W are the height and width of different level feature blocks. In the transformation branch, the sequence of three operations is used to generate robust and meaningful saliency features, as is shown in Figure 2. The detailed parameters of the three different operators can be found in Table 1. After the processes above, we obtain five different feature blocks $\{T^i \in R^{H \times W \times C} | i = 1, \ldots, 5\}$, which are to be used in the decoder part of the network.

3.3. Decoder Module. In the encoder, different levels of feature blocks contain different information. The high-layer feature blocks encode the semantic information for category, and these layers do not care about local detail for image.

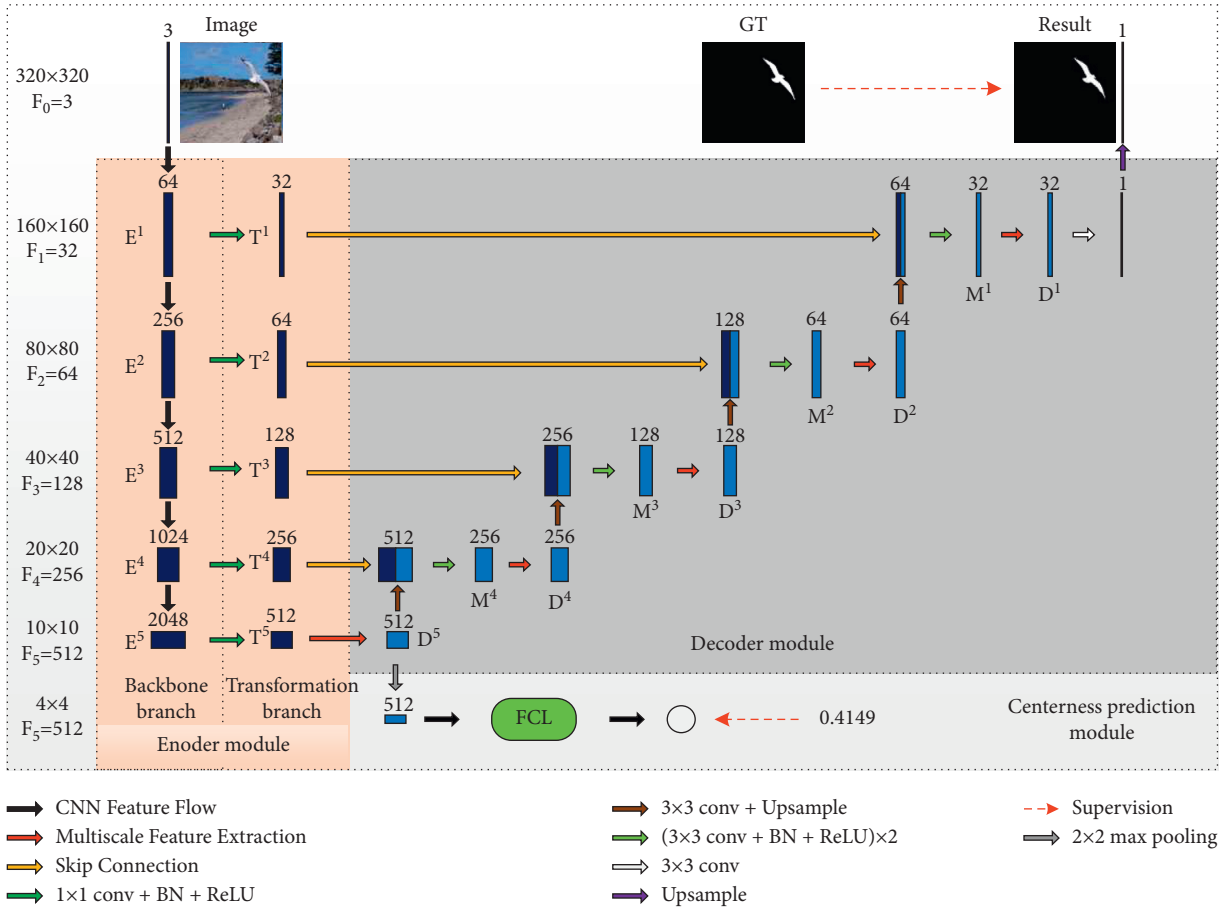


Figure 2: The overall framework of our proposed multiscale deep network (M2Net). M2Net is based on ResNet-50 [45] or ResNeXt-101 [46] with supervision from saliency map and object position. M2Net consists of the backbone branch, transformation branch, decoder branch, and centerness prediction branch. The backbone network is used to extract some important saliency features, the transformation part is used to generate robust saliency features, the decoder part is used to generate the final saliency maps, and the last part is used to predict object position in an image.

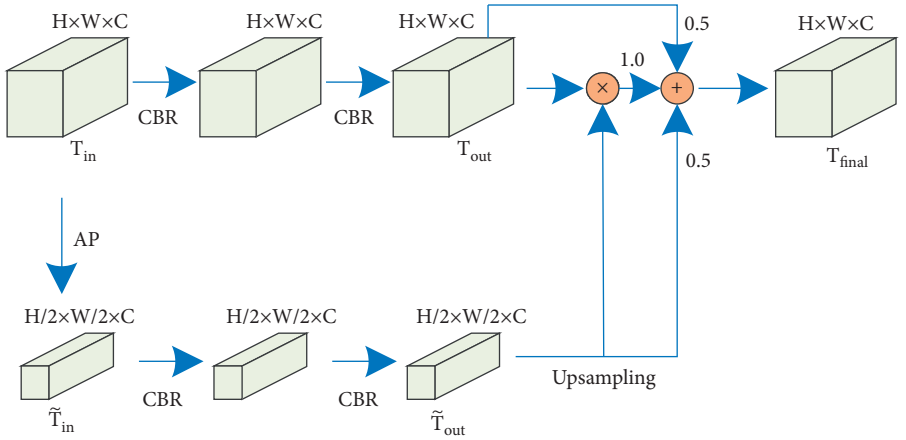


Figure 3: Multiscale feature extraction unit; CBR is composed of convolution, batch normalization, and ReLU operation; AP denotes average pooling.

The low-layer feature blocks contain more detailed information about the image, and these layers suffer from the problem of semantic ambiguity. The decoder module is designed to integrate these different feature blocks. The combination of these different level feature blocks can enhance the representation ability to complete the saliency

Table 1: Detailed parameters of three different operators, including convolutional operator, batch normalization operator, and ReLU operator, respectively. The left two parameters of Conv2d are the number of input and output channels, and the right two parameters are the kernel size. The parameter of the BatchNorm2d operator is the number of feature channels, and the parameter of the ReLU operator is in place.

No.	Feature	Conv2d	BatchNorm2d	ReLU	Result
1	F^1	2048, 512, 1×1	512	True	T^1
2	F^2	1024, 256, 1×1	256	True	T^2
3	F^3	512, 128, 1×1	128	True	T^3
4	F^4	256, 64, 1×1	64	True	T^4
5	F^5	128, 32, 1×1	32	True	T^5

prediction task. The decoder network comprises two main computing units: (i) Multiscale feature extraction unit, which can extract multiscale contextual features to facilitate saliency prediction models to extract discriminative features. (ii) Cross-layer feature block fusion unit, which continually fuses adjacent level feature blocks from $\{T^i|i=1,\ldots,5\}$.

Figure 3 shows the details of the multiscale feature extraction unit. Given a feature block $F^i_{\text{in}} \in R^{H\times W\times C}$, we first use average pooling to perform a downsampling operation. After that, we can obtain $\widetilde{F}^i_{\text{in}} \in R^{(H/2)\times(W/2)\times C}$. To obtain robust saliency features, the two branches are processed by combination operation, which is composed of convolution, batch normalization, and ReLU operation. After a series of above-mentioned processing, we obtain $F^i_{\text{out}} = \text{CBR}(\text{CBR}(F^i_{\text{in}}))$ and $\widetilde{F}^i_{\text{out}} = \text{CBR}(\text{CBR}(\widetilde{F}^i_{\text{in}}))$. The output of the bottom branch $\widetilde{F}^i_{\text{out}}$ is upsampled to match the output of top branch F^i_{out}. To extract the multiscale features, we integrate the two branches by using multiplication and addition operations. The multiscale feature extraction unit is formulated as follows:

$$F^i_{\text{final}} = 0.5\times F^i_{\text{out}} + 0.5\times \text{Up}\left(\widetilde{F}^i_{\text{out}}\right) + F^i_{\text{out}}\times \text{Up}\left(\widetilde{F}^i_{\text{out}}\right), \quad (1)$$

where $U(\cdot)$ denote upsampling operation.

Figure 4 shows the details of the cross-layer feature block fusion unit. This transformation unit is composed of four different kinds of operators, including convolution, upsampling, concatenation, and combinator. The first convolution layer can halve the number of channels for high-level feature block. To adapt to the low-level feature block, the transformed high-level feature block is processed by the second upsampling operator, which can increase the size of feature blocks by 2 times. Then, the concatenation operator is used to build one larger feature block. Finally, the combinator is composed of the two repeated cascaded structures of convolution operators, each of them followed by a batch normalization layer and a ReLU layer. After a series of above-mentioned processing, we can obtain four different feature blocks $\{M^i \in R^{H\times W\times C}|i=1,\ldots,4\}$, as shown in Figure 2.

The cross-layer feature block fusion unit is formulated as follows:

$$M^i = \text{CBR}\left(\text{Cat}\left(T^i, U\left(D^{i+1}\right)\right)\right), \quad (2)$$

where $\text{CBR}(\cdot)$ and $U(\cdot)$ represent the combined operation as mentioned above and the upsampling operation, respectively. D^i denote the output of the decoder, and it is formulated as follows:

$$D^i = \begin{cases} \text{MSF}\left(T^5\right), & i=5, \\ \text{MSF}\left(M^i\right), & i=1,2,3,4, \end{cases} \quad (3)$$

where $\text{MSF}(\cdot)$ denote multiscale feature extraction operation, which is defined in equation (1).

3.4. Centerness Prediction Module. Object location information can be very useful to improve the image classification task but seldom used in the saliency detection task.

In this paper, we introduce the centerness to the saliency detection. We define centerness as a ratio between EO and EC, as is shown in Figure 5. The location of node O represents the center of ground truth or saliency map, and it can be calculated as follows:

$$\left(O_x, O_y\right) = \left(\frac{\sum_{i=1}^{H}\sum_{j=1}^{W} x_i f_{ij}}{\sum_{i=1}^{H}\sum_{j=1}^{W} f_{ij}}, \frac{\sum_{i=1}^{H}\sum_{j=1}^{W} y_i f_{ij}}{\sum_{i=1}^{H}\sum_{j=1}^{W} f_{ij}}\right), \quad (4)$$

where $f_{ij} \in [0,1]$ denotes the gray value of ground truth or predicted saliency map.

Centerness prediction module comprised two main components: (i) FCL (two fully connected layers), which directly maps high-dimensional feature space to 1-dimensional feature space; (ii) logistic function, which applies a sigmoid function to restrict the number from a large scale to within the range 0–1. Figure 6 shows the details of the FCL. This component contains three fully connected layers, and each layer contains a different number of neural nodes.

3.5. Deep Supervision. An effective loss function plays an important role in training more effective saliency models [26]. When the image contains complex background, the deep network with BCE loss will probably generate poor saliency results. To generate high-quality saliency maps with clear boundaries, we propose a centerness-aware loss, which is defined as follows:

$$L_{\text{CAL}} = L_{\text{bce}} + L_{\text{loc}} + \lambda\times L_{\text{iou}}, \quad (5)$$

where L_{bce}, L_{loc}, and L_{iou} denote BCE loss [49], location loss, and IoU loss [17, 27], respectively. The parameter λ is a hyperparameter which is set to 0.5 in this paper.

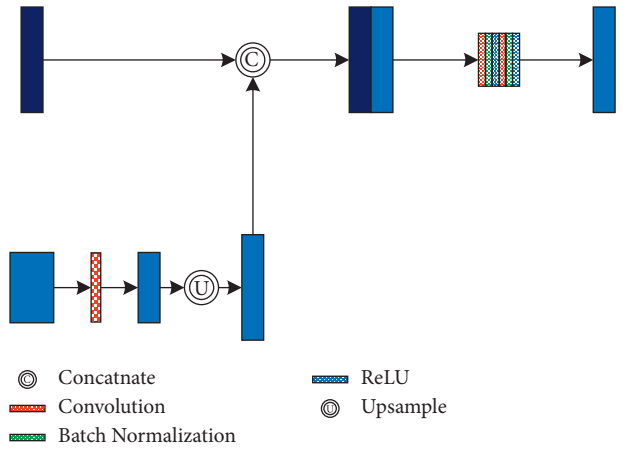


Figure 4: Cross-layer feature block fusion unit, which continually fuses adjacent level feature blocks.

Binary cross entropy (BCE) is a widely used loss in saliency detection tasks, and it is defined as follows:

$$L_{\text{bce}} = -\frac{1}{B \times M} \sum_{k=0}^{k=B} \sum_{i=0}^{i=M} \left[y_i^{(k)} \times \log x_i^{(k)} + \left(1 - y_i^{(k)}\right) \times \log\left(1 - x_i^{(k)}\right) \right], \tag{6}$$

where x_i and $y_i \in \{0, 1\}$ denote the prediction of the pixel i and ground truth. B is the batch size and M is the product of the height and width of a given image.

The position of salient objects is very important information. Hence, we introduced it into our training loss; L_{loc} is defined as follows:

$$L_{\text{loc}} = \frac{1}{B} \sum_{k=0}^{k=B} \left(C_g^k - C_p^k\right)^2, \tag{7}$$

where C_g^k denotes the ground truth of the k-th image and C_p^k is the result of centerness prediction.

To uniformly highlight the whole salient region, we integrated IoU [17, 27] into our training loss. It is defined as

$$L_{\text{iou}} = \frac{1}{B} \sum_{k=1}^{B} \left(1 - \frac{\sum_{i=0}^{M} s_i^k y_i^k}{\sum_{i=0}^{M} \left[s_i^k + y_i^k - s_i^k y_i^k \right]} \right). \tag{8}$$

where $s_i^k \in \{0, 1\}$ is the predicted probability of being the foreground object and y_i^k is the ground truth of the pixel i.

4. Experiments

4.1. Implementation Details. We train our saliency model on the DUTS-TR [22] dataset with 10553 images as followed by [3, 16]. For a fair comparison, we use ResNet and ResNeXt as backbone networks, respectively. For convenience, all the training and testing images are resized to 320×320. Our saliency model is implemented in PyTorch. The parameters of backbone networks are initialized with the models pretrained on the classification dataset. All the other parameters of M2Net are set by the default setting of PyTorch 1.2.0. The hyperparameters are set as follows: weight decay = 0.0005 and momentum = 0.9, and the initial learning rate is set to 0.005 for pretrained backbone networks [45, 46] and 0.05 for the rest parts. In this paper, we use the warm-up and linear decay methods to dynamically adjust the learning rate. During the training stage, random flip, random contrast, random saturation, and random brightness act as data augmentation techniques to avoid the overfitting problem. We apply a stochastic gradient descent algorithm to update all the parameters of the proposed M2Net. To ensure model convergence, M2Net is trained for 32 epochs with a mini-batch of 15 on an NVIDIA GTX 2080 Ti GPU.

4.2. Datasets. The performance of M2Net is evaluated on five benchmark saliency datasets, including ECSSD [50], PASCAL-S [51], DUTS [22], DUT-OMRON [30], and HKU-IS [34]. ECSSD [50] contains 1000 meaningful semantic images with pixel-accurate annotations. The PASCAL-S [51] dataset is composed of 850 challenging images, which are carefully selected from the PASCAL VOC segmentation dataset. DUTS is the largest salient object detection (SOD) dataset. It contains 10553 images for training and the remaining 5019 images for testing. DUT-OMRON [30] is composed of 5168 high-quality but challenging images. Images in this dataset contain one or more salient objects with complex background. The HKU-IS [34] contains 4447 challenging images which have multiple disconnected salient objects.

4.3. Evaluation Criteria. To quantitatively evaluate the performance, four measurements, including Precision-Recall (PR) curve, F-measure, and Mean Absolute Error (MAE), and S-measure, are adopted in our experiments.

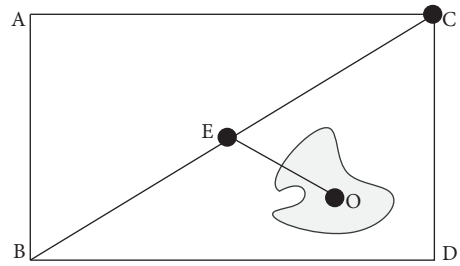


Figure 5: Illustration of the centerness calculation. We define centerness as the ratio between EO and EC.

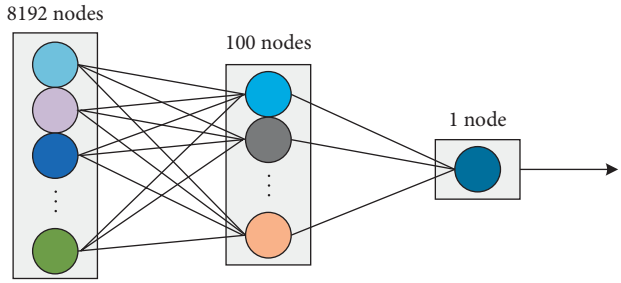


Figure 6: Illustration of the detailed FCL (fully connected layer). The first column contains 8192 nodes, the second column contains 100 nodes, and the third column contains only one node.

Precision-Recall curve is a widely used graphical tool to evaluate the robustness of saliency maps. It can demonstrate the relation of precision and recall by thresholding the final saliency results from 0 to 255. The larger the area under the PR curve, the better the performance.

The F-measure is a weighted combination of precision and recall, which is defined as

$$F_\beta = \frac{\left(1+\beta^2\right) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}, \tag{9}$$

where β^2 is set to 0.3 as done in most recent state-of-the-art saliency methods [3, 16, 19, 52–56] to emphasize the precision. The mean F-measure (F_m) of each benchmark dataset is reported in the paper.

Mean Absolute Error (MAE) is a metric, which measures the pixelwise average absolute difference between saliency map and its corresponding ground truth. The MAE score is defined as follows:

$$\text{MAE} = \frac{1}{M} \sum_{m=1}^{M} |x_m - y_m|, \tag{10}$$

where x and y are the prediction result and ground truth, respectively, and M indicates the total number of image pixels.

S-measure is more sensitive to foreground structural information of saliency maps, which is closer to the human visual system. It considers the object-aware structural similarity S_o and the region-aware structural similarity S_r:

$$S = \gamma \times S_o + (1-\gamma) \times S_r, \tag{11}$$

where γ is set to 0.5 as suggested in [3, 20, 53, 57].

4.4. Comparison with State of the Art. In this section, we compare our method with seventeen previous state-of-the-art saliency models, including NLDF [58], Amulet [15], R3Net [59], RAS [60], DGRL [61], C2SNet [54], PiCANet [16], BMPM [43], BASNet [17], AFNet [62], SCRN [63], CPD [64], EGNet [20], PoolNet [18], F3Net [19], MINet [53], and GateNet [3]. Note that all the saliency maps of above saliency methods are produced by running source codes or precomputed by the authors.

4.4.1. Quantitative Evaluation. To fully compare the proposed saliency model with these state-of-the-art methods, the detailed experimental results in terms of three metrics are listed in Table 2. For better comparison, we use the ResNet-50 and ResNeXt-101 as backbone networks for training our proposed M2Net. Specifically, our method achieves a great improvement in terms of the F_m compared to the most recent saliency model GateNet [3] on the challenging DUT-TE [22] (0.857 versus 0.816), DUT-OMRON [30] (0.791 versus 0.762), and PASCAL-S [51] (0.858 versus 0.827). In addition, we demonstrate the standard PR curves in Figures 7 and 8. Our method achieves the best performance on the ECSSD, HKU-IS, PASCAL-S, DUT-OMRON, and DUT-TE datasets.

4.4.2. Qualitative Evaluation. Some prediction results of the proposed M2Net and ten state-of-the-art saliency methods have been shown in Figure 9. We observe that the proposed method M2Net not only uniformly highlights the correct salient object region clearly but also well suppresses the background clutter effectively. It excels in dealing with various challenging scenarios, including small objects (rows

Table 2: Performance comparison with state-of-the-art methods on five popular saliency datasets. MAE (smaller is better), max F-measure (larger is better), and E-measure (larger is better) are used to measure the model performance.

Method	DUT-OMRON			DUTS			ECSSD			PASCAL-S			HKU-IS		
	F_m	MAE	S_m	F_m	MAE	S_m	F_m	MAE	S_m	F_m	MAE	S_m	F_m	MAE	S_m
VGG backbone															
$Amulet_{17}$	0.647	0.098	0.781	0.678	0.085	0.804	0.868	0.059	0.894	0.769	0.099	0.819	0.841	0.051	0.886
$NLDF_{17}$	0.684	0.080	0.770	—	—	—	0.878	0.063	0.875	0.780	0.101	0.801	0.874	0.048	0.879
$BMPM_{18}$	0.692	0.064	0.809	0.745	0.049	0.862	0.868	0.045	0.911	0.771	0.075	0.845	0.871	0.039	0.907
$C2SNet_{18}$	0.683	0.072	0.798	0.716	0.063	0.828	0.864	0.055	0.893	0.769	0.083	0.835	0.851	0.048	0.883
RAS_{18}	0.713	0.062	0.814	0.751	0.059	0.839	0.889	0.056	0.893	0.785	0.106	0.793	0.871	0.045	0.887
$PiCANet_{18}$	0.710	0.068	0.826	0.749	0.054	0.861	0.885	0.046	0.914	0.801	0.079	0.849	0.870	0.042	0.906
CPD_{19}	0.745	0.057	0.818	0.813	0.043	0.867	0.915	0.040	0.910	0.830	0.075	0.841	0.896	0.033	0.904
$MINet_{20}$	0.741	0.057	0.821	0.823	0.039	0.875	0.922	0.036	0.919	0.840	0.066	0.852	0.904	0.031	0.912
ResNet backbone															
$DGRL_{18}$	0.733	0.062	0.806	0.794	0.050	0.842	0.906	0.041	0.903	0.827	0.073	0.837	0.890	0.036	0.894
$PiCANet_{18}$	0.717	0.065	0.832	0.759	0.051	0.869	0.886	0.046	0.917	0.802	0.078	0.852	0.870	0.043	0.904
$BASNet_{19}$	0.756	0.057	0.836	0.791	0.048	0.866	0.880	0.037	0.916	0.777	0.079	0.834	0.896	0.032	0.909
$EGNet_{19}$	0.756	0.053	0.841	0.815	0.039	0.887	0.920	0.037	0.925	0.829	0.076	0.850	0.901	0.031	0.918
$PoolNet_{19}$	0.747	0.056	0.836	0.809	0.040	0.883	0.915	0.039	0.921	0.828	0.076	0.849	0.899	0.032	0.917
CPD_{19}	0.747	0.056	0.825	0.805	0.043	0.869	0.917	0.037	0.918	0.829	0.074	0.844	0.891	0.034	0.906
$SCRN_{19}$	0.746	0.056	0.837	0.809	0.040	0.885	0.918	0.038	0.927	0.837	0.066	*0.865*	0.897	0.034	0.916
$MINet_{20}$	0.755	0.055	0.833	0.828	0.037	0.884	0.925	0.034	0.925	0.840	0.066	0.854	0.909	*0.029*	0.919
$GateNet_{20}$	0.746	0.055	0.838	0.807	0.040	0.885	0.916	0.040	0.920	0.830	0.071	0.854	0.899	0.033	0.915
Ours	*0.774*	0.054	0.846	*0.837*	0.038	0.889	*0.926*	*0.033*	0.927	*0.847*	*0.064*	0.863	*0.913*	*0.029*	0.922
ResNeXt backbone															
$R3Net_{18}$	0.747	0.062	0.815	—	—	—	0.914	0.040	0.910	0.803	0.095	0.803	0.894	0.036	0.895
$GateNet_{20}$	0.762	*0.051*	*0.849*	0.816	*0.035*	*0.897*	0.917	0.035	0.929	0.827	0.065	*0.865*	0.903	0.030	*0.925*
Ours	**0.791**	**0.049**	**0.860**	**0.857**	**0.034**	**0.900**	**0.934**	**0.032**	**0.933**	**0.858**	**0.061**	**0.870**	**0.923**	**0.025**	**0.929**

Bold, italics, and underline indicate the best, second best, and third best performance. "—" means that the author has not provided corresponding saliency maps.

2, 8, and 9), cluttered backgrounds (rows 1, 3, and 5), low contrast between the salient object and background region (rows 4 and 7), and image boundary (row 6). Compared with other state-of-the-art methods, the detected object boundaries of our saliency map are clear and sharper. Most importantly, the proposed saliency model M2Net achieves these results without any postprocessing.

4.5. Ablation Analysis. Before analyzing the influence of each saliency module, there is one hyperparameter λ to be determined. λ is used in CAL loss to balance different losses. Table 3 lists the scores of F_m, MAE, and S_m when λ gives four discrete values. As can be seen, when λ equals 0.50, these indicators reach the best results. To investigate the importance of different components in our proposed M2Net, we will conduct a detailed analysis next.

4.5.1. Effectiveness of Backbones. In the saliency detection, VGG [65], ResNet [45], and ResNeXt [46] are widely used as the pretrained backbone. Table 2 demonstrates that ResNet-50 and ResNeXt-101 can achieve better performance compared with VGG in most cases. To demonstrate the effectiveness of ResNet-50 and ResNeXt-101, we also selected two widely used datasets for evaluation, and the comparison results are shown in Table 4. From Table 4, we can see that M2Net with ResNeXt-101 [46] can get better performance compared with ResNet-50 [45].

4.5.2. Effectiveness of Components. We take an FPN-like network as our baseline network. Then, we install the multiscale feature unit on the baseline network and evaluate its performance. The comparison results are shown in Table 4. It can be seen that a multiscale feature unit can achieve significant improvement over the FPN-like network. We also quantitatively evaluate the effect of the centerness-aware loss in Table 4. Compared to "+B," the proposed M2Net with the CAL achieves consistent performance enhancements in terms of three metrics. Visual comparison of saliency maps generated by BCE loss and our centerness-aware loss are shown in Figures 10(c), 10(d), and 10(e). To fully compare CAL and three other losses, including FLoss, PPA, and IoU loss, the detailed experimental results are listed in Table 5. As it can be seen, our proposed CAL loss can get the best results on two challenge saliency datasets.

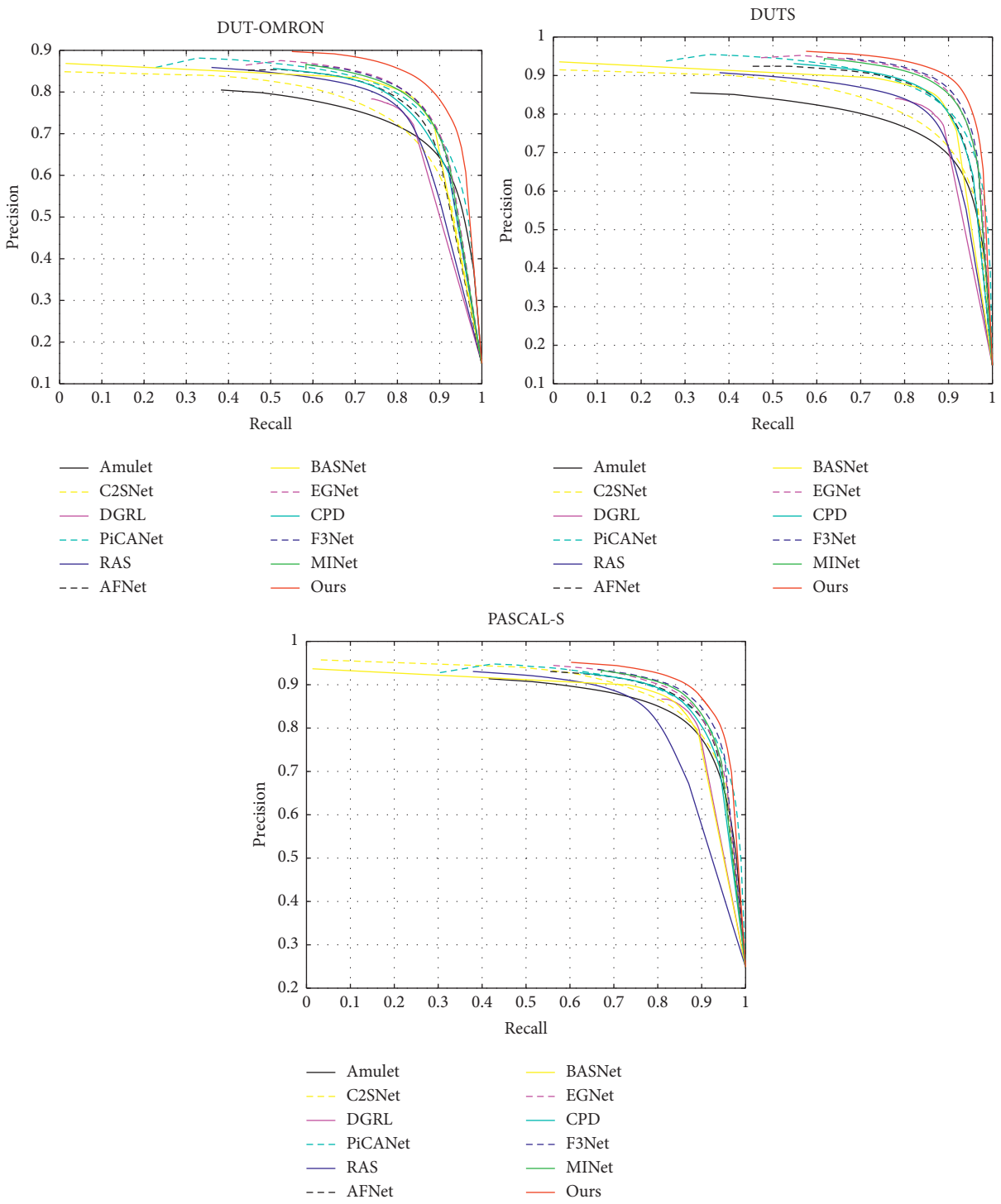


Figure 7: Precision-Recall curves on three saliency datasets, including DUT-OMRON, DUTS, and PASCAL-S.

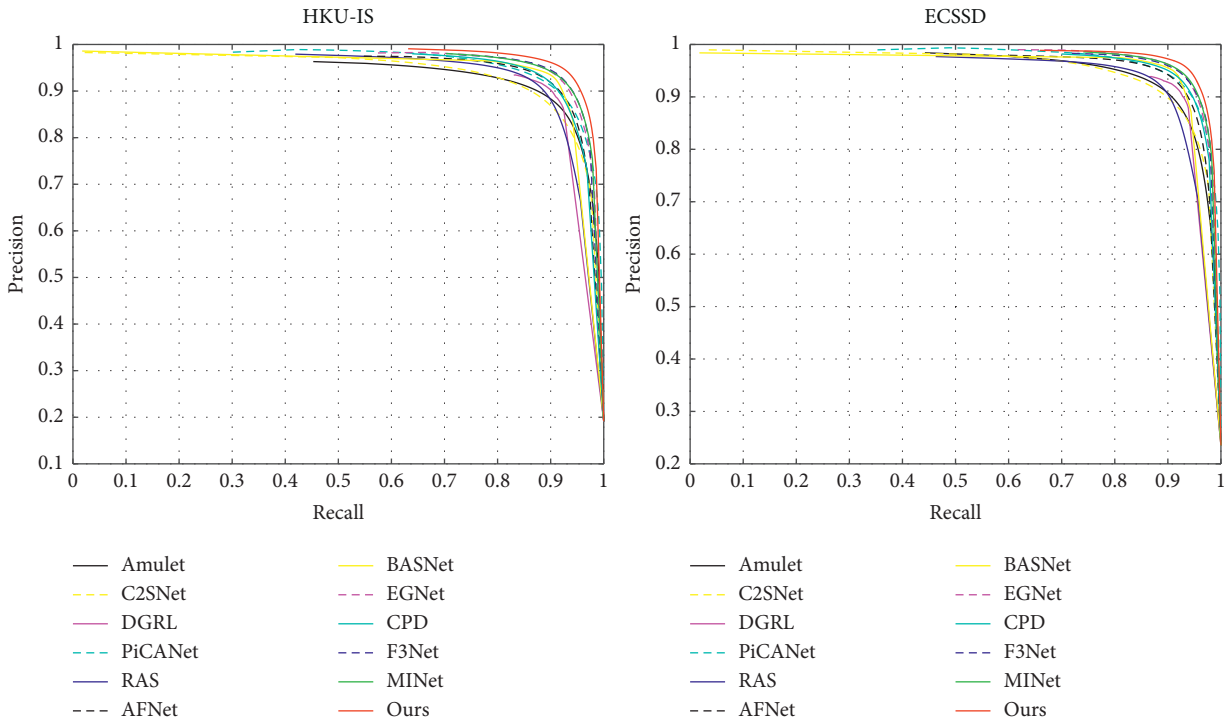


Figure 8: Precision-Recall curves on two common saliency datasets, including HKU-IS and ECSSD.

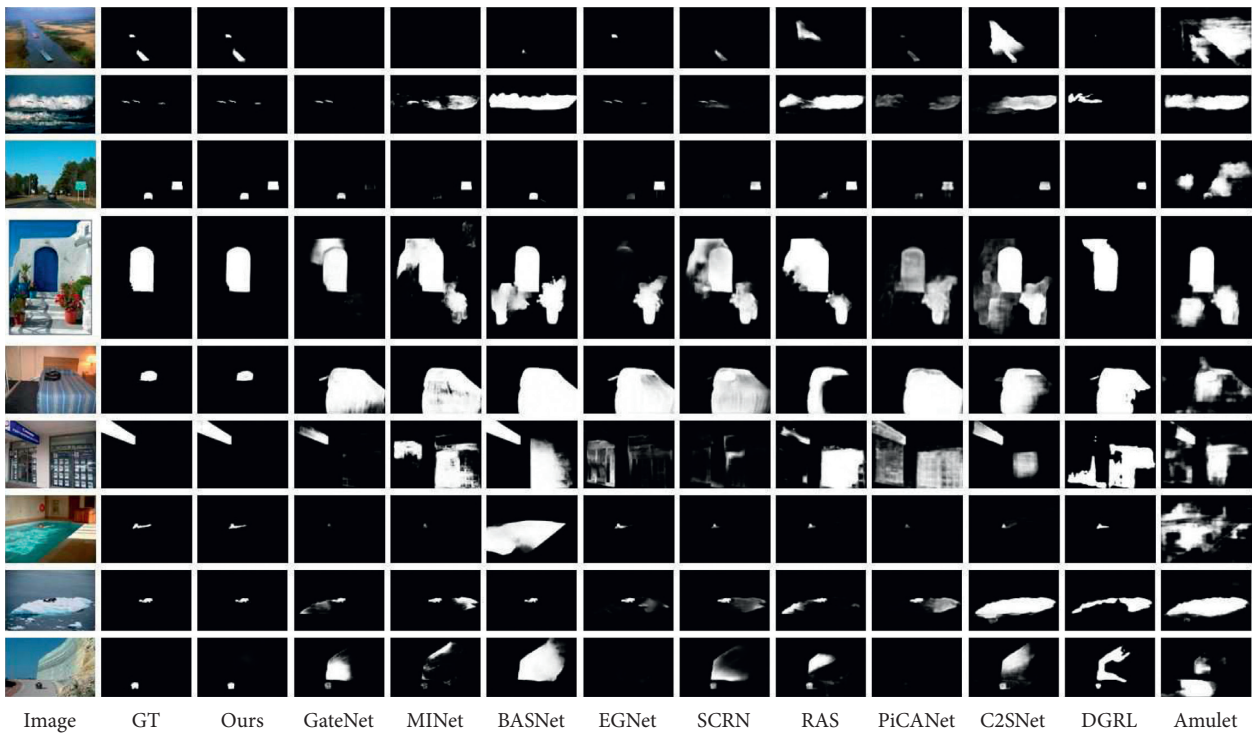


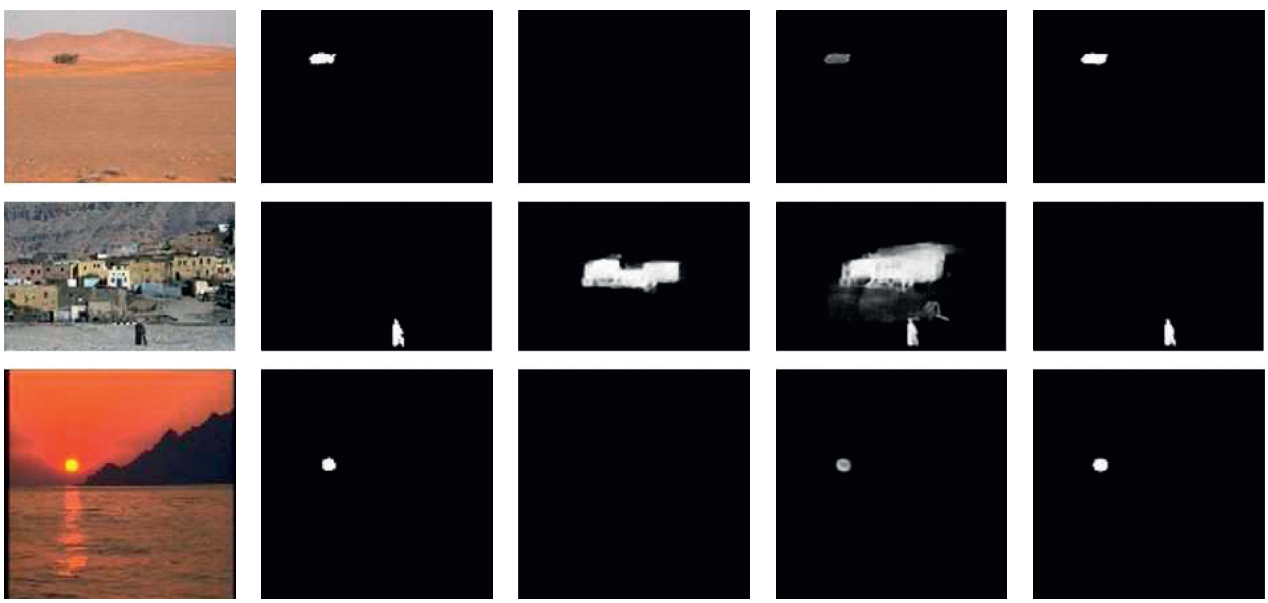
Figure 9: Visual comparison of our method with ten state-of-the-art methods.

Table 3: Comparison with different λ. When $\lambda = 0.50$, the proposed model achieves the best results.

	DUT-OMRON			DUTS			ECSSD			PASCAL-S			HKU-IS		
	F_m	MAE	S_m	F_m	MAE	S_m	F_m	MAE	S_m	F_m	MAE	S_m	F_m	MAE	S_m
$\lambda = 0.25$	0.761	0.054	0.846	0.825	0.038	0.892	0.918	0.036	0.928	0.837	0.067	0.862	0.907	0.030	0.923
$\lambda = 0.50$	0.791	0.049	0.860	0.857	0.034	0.900	0.934	0.032	0.933	0.858	0.061	0.870	0.923	0.025	0.929
$\lambda = 0.75$	0.777	0.051	0.847	0.839	0.038	0.887	0.926	0.034	0.926	0.849	0.065	0.861	0.913	0.029	0.920
$\lambda = 1.00$	0.778	0.051	0.845	0.844	0.036	0.887	0.927	0.035	0.924	0.845	0.069	0.853	0.917	0.028	0.919

Table 4: Ablation analysis on two challenge datasets. FPN-1: FPN [48] with ResNet-50 backbone; FPN-2: FPN [48] with ResNeXt-101 backbone; M2N-1: the proposed M2Net with ResNet-based backbone; M2N-2: the proposed M2Net with ResNeXt-based backbone B: binary cross entropy; C: centerness; I: Intersection over Union.

Model	Loss	DUT-OMRON			PASCAL-S		
		F_m	MAE	S_m	F_m	MAE	S_m
FPN-1	+B	0.711	0.064	0.815	0.814	0.074	0.849
M2N-1	+B	0.742	0.055	0.840	0.830	0.069	0.861
M2N-1	+B + C	0.748	0.055	0.842	0.836	0.067	0.862
M2N-1	+B + C + I	0.774	0.054	0.846	0.847	0.064	0.863
FPN-2	+B	0.745	0.056	0.842	0.835	0.067	0.861
M2N-2	+B	0.761	0.051	0.851	0.845	0.063	0.864
M2N-2	+B + C	0.768	0.050	0.855	0.852	0.061	0.866
M2N-2	+B + C + I	0.791	0.049	0.860	0.858	0.061	0.870



(a) (b) (c) (d) (e)

Figure 10: Visual comparisons for showing the proposed modules. M: the proposed M2Net with binary cross entropy loss; C: centerness; I: Intersection over Union. (a) Image. (b) GT. (c) M. (d) M + C. (e) M + C + I.

Table 5: Comparison of the proposed loss and three other methods on two challenge datasets. M2N: the proposed M2Net with ResNeXt-based backbone CAL: centerness-aware loss; I: Intersection over Union loss [17, 27]; FLoss: F-measure based loss [28]; PPA: pixel position aware loss [19].

Model	DUT-OMRON			PASCAL-S		
	F_m	MAE	S_m	F_m	MAE	S_m
M2N + I	0.750	0.048	0.802	0.813	0.078	0.813
M2N + FLoss	0.781	0.051	0.817	0.850	0.074	0.829
M2N + PPA	0.770	0.054	0.841	0.845	0.066	0.857
M2N + CAL	0.791	0.049	0.860	0.858	0.061	0.870

5. Conclusion

In this paper, we proposed a multiscale deep network with a centerness-aware loss for salient object detection. The proposed M2Net aims to solve saliency prediction and centerness prediction simultaneously. Our model consists of a bottom-up encoder module, top-down decoder module, and centerness prediction module. In the encoder, we use the pretrained backbone to extract multilevel saliency features from preprocessed images. Next, in the decoder module, we use a skip/concatenation connection scheme. To generate the final saliency maps, we proposed a novel multiscale feature extraction method. Lastly, we design a centerness prediction module, which can help to uniformly highlight the entire salient object. Extensive experimental results on five widely used datasets demonstrate that our method outperforms 17 state-of-the-art approaches under different evaluation metrics.

Data Availability

The labeled datasets used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of Shandong Province (nos. ZR2019PF019 and ZR2020QF044).

References

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[2] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Ma- Chine Intelligence*, vol. 37, no. 3, pp. 569–582, 2014.

[3] X. Zhao, Y. Pang, L. Zhang, H. Lu, and L. Zhang, "Suppress and balance: a simple gated network for salient object detection," in *Proceedings of the European Conference on Computer Vision*, Glasgow, UK, August 2020.

[4] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?" in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, Washington, DC, USA, June 2004.

[5] X. Qin, S. He, C. Perez Quintero, A. Singh, M. Dehghan, and J. Martin, "Real-time salient closed boundary tracking via line segments perceptual grouping," in *Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4284–4289, IEEE, Vancouver, BC, Canada, September 2017.

[6] J. He, J. Feng, X. Liu et al., "Mobile product search with bag of hash bits and boundary reranking," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3005–3012, IEEE, Providence, RI, USA, June 2012.

[7] X. Wang, X. Liang, B. Yang, and F. W. B. Li, "No-reference synthetic image quality assessment with convolutional neural network and local image saliency," *Computational Visual Media*, vol. 5, no. 2, pp. 193–208, 2019.

[8] C. Craye, D. Filliat, and J.-F. Goudou, "Environment exploration for object-based visual saliency learning," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2303–2309, IEEE, Stockholm, Sweden, May 2016.

[9] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1304–1318, 2004.

[10] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 185–198, 2009.

[11] J.-Y. Zhu, J. Wu, Y. Xu, E. Chang, and Z. Tu, "Unsupervised object class discovery via saliency-guided multiple class learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 4, pp. 862–875, 2014.

[12] A. Karpathy, S. Miller, and Li Fei-Fei, "Object discovery in 3D scenes via shape analysis," in *Proceedings of the 2013 IEEE International Conference on Robotics and Automation*, pp. 2088–2095, IEEE, Karlsruhe, Germany, May 2013.

[13] S. Frintrop, G. M. Garc´ıa, and B. Cremers, "A cognitive approach for object discovery," in *Proceedings of the 2014 22nd International Conference on Pattern Recognition*, pp. 2329–2334, IEEE, Stockholm, Sweden, August 2014.

[14] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, Boston, MA, USA, June 2015.

[15] P. Zhang, D. Wang, H. Lu, H. Wang, and R. Xiang, "Amulet: aggregating multi-level convolutional features for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 202–211, Venice, Italy, October 2017.

[16] N. Liu, J. Han, and M.-H. Yang, "Picanet: learning pixel-wise contextual attention for saliency detection," in *Proceedings of*

the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3089–3098, Salt Lake City, UT, USA, June 2018.

[17] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and J. Martin, "Basnet: boundary-aware salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7479–7489, Long Beach, CA, USA, June 2019.

[18] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3917–3926, Long Beach, CA, USA, June 2019.

[19] J. Wei, S. Wang, and Q. Huang, "F^3Net: fusion, feedback and focus for salient object detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 12321–12328, 2020.

[20] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "Egnet: edge guidance network for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8779–8788, Seoul, South Korea, October 2019.

[21] H. Zhou, X. Xie, J.-H. Lai, Z. Chen, and L. Yang, "Interactive two-stream decoder for ac- curate and fast saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9141–9150, Seattle, WA, USA, June 2020.

[22] L. Wang, H. Lu, Y. Wang et al., "Learning to detect salient objects with image-level supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 136–145, Honolulu, HI, USA, July 2017.

[23] G. Li, X. Yuan, and L. Lin, "Weakly supervised salient object detection using image labels," in *Proceedings of the AAAI Conference on Artificial Intelligence*, New Orleans, LA, USA, February 2018.

[24] J. Zhang, T. Zhang, Y. Dai, M. Harandi, and R. Hartley, "Deep unsupervised saliency detection: a multiple noisy labeling perspective," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9029–9038, Salt Lake City, UT, USA, June 2018.

[25] G. Li, X. Yuan, L. Lin, and Y. Yu, "Instance- level salient object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2386–2395, Honolulu, HI, USA, July 2017.

[26] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: an in-depth survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[27] M. A. Rahman and Y. Wang, "Optimizing intersection-over-union in deep neural networks for image segmentation," in *Proceedings of the International Symposium on Visual Computing*, pp. 234–244, Springer, Las Vegas, NV, USA, December 2016.

[28] K. Zhao, S. Gao, W. Wang, and M.-M. Cheng, "Optimizing the *F*-measure for threshold-free salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8849–8857, Seoul, South Korea, October 2019.

[29] Z. Jiang and L. S. Davis, "Submodular salient region detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2043–2050, Portland, OR, USA, June 2013.

[30] C. Yang, L. Zhang, H. Lu, R. Xiang, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3166–3173, Portland, OR, USA, June 2013.

[31] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1265–1274, Boston, MA, USA, June 2015.

[32] S. He, R. W. H. Lau, W. Liu, Z. Huang, and Q. Yang, "SuperCNN: a superpixelwise convolutional neural network for salient object detection," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 330–344, 2015.

[33] G. Lee, Y-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 660–668, Las Vegas, NV, USA, June 2016.

[34] G. Li and Y. Yu, "Visual saliency based on multi- scale deep features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5455–5463, Boston, MA, June 2015.

[35] L. Wang, H. Lu, R. Xiang, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3183–3192, Boston, MA, USA, June 2015.

[36] J. Kim and V. Pavlovic, "A shape-based approach for salient object detection using deep learning," in *Proceedings of the European Conference on Computer Vision*, pp. 455–470, Springer, Amsterdam, Netherlands, October 2016.

[37] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech, "Unconstrained salient object detection via proposal subset optimization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5733–5742, Las Vegas, NV, USA, June 2016.

[38] J. Zhang, J. Xie, and N. Barnes, "Learning noise-aware encoder-decoder from noisy labels by alternating back-propagation for saliency detection," in *Proceedings of the European Conference on Computer Vision*, pp. 349–366, Springer, July 2020.

[39] T. Zhao and X. Wu, "Pyramid feature attention net- work for saliency detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3085–3094, Long Beach, CA, USA, June 2019.

[40] S. Mohammadi, M. Noori, A. Bahri, S. Ghofrani Majelan, and M. Havaei, "Cagnet: content-aware guidance for salient object detection," *Pattern Recognition*, vol. 103, Article ID 107303, 2020.

[41] L.-C. Chen, P. George, I. Kokkinos, K. Murphy, and L. Alan, "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, 2017.

[42] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2881–2890, Honolulu, HI, USA, July 2017.

[43] L. Zhang, J. Dai, H. Lu, H. You, and G. Wang, "A bi-directional message passing model for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1741–1750, Salt Lake City, UT, USA, June 2018.

[44] W. Wang, S. Zhao, J. Shen, S. C. H. Hoi, and B. Ali, "Salient object detection with pyramid attention and salient edges," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1448–1457, Long Beach, CA, USA, June 2019.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference*

on Computer Vision and Pattern Recognition, pp. 770–778, Las Vegas, NV, USA, June 2016.

[46] S. Xie, R. Girshick, P. Dolla´r, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1492–1500, Honolulu, HI, USA, July 2017.

[47] S. S. Kruthiventi, V. Gudisa, J. H. Dholakiya, and R. Venkatesh Babu, "Saliency unified: a deep architecture for simultaneous eye fixation prediction and salient object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5781–5790, Las Vegas, NV, USA, June 2016.

[48] T.-Yi Lin, P. Dolla´r, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125, Honolulu, HI, USA, July 2017.

[49] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Annals of Operations Research*, vol. 134, no. 1, pp. 19–67, 2005.

[50] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1155–1162, Portland, OR, USA, June 2013.

[51] L. Yin, X. Hou, C. Koch, M. R. James, and L. Y. Alan, "The secrets of salient object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 280–287, Columbus, OH, USA, June 2014.

[52] Q. Hou, M.-M. Cheng, X. Hu, B. Ali, Z. Tu, and P. H. S. Torr, "Deeply supervised salient object detection with short connections," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3203–3212, Honolulu, HI, USA, July 2017.

[53] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9413–9422, Seattle, WA, USA, June 2020.

[54] X. Li, Y. Fan, H. Cheng, W. Liu, and D. Shen, "Contour knowledge transfer for salient object detection," in *Proceedings of the European Conference on Computer Vision*, pp. 355–370, Munich, Germany, October 2018.

[55] M. Ma, C. Xia, and L. Jia, "Pyramidal feature shrinking for salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2311–2318, Palo Alto, California USA, February 2021.

[56] J. Wei, S. Wang, Z. Wu, C. Su, Q. Huang, and T. Qi, "Label decoupling framework for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 13025–13034, Seattle, WA, USA, June 2020.

[57] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and B. Ali, "Structure-measure: a new way to evaluate foreground maps," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4548–4557, Venice, Italy, October 2017.

[58] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6609–6617, Honolulu, HI, USA, July 2017.

[59] Z. Deng, X. Hu, L. Zhu et al., "R3net: recurrent residual refinement network for saliency detection," in *Proceedings of the Twenty-Seventh International Joint Conference on AAAI*, pp. 684–690, Stockholm Sweden, July 2018.

[60] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *Proceedings of the European Conference on Computer Vision*, pp. 234–250, Munich, Germany, October 2018.

[61] T. Wang, L. Zhang, S. Wang et al., "Detect globally, refine locally: a novel approach to saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3127–3135, Salt Lake City, UT, USA, June 2018.

[62] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1623–1632, Long Beach, CA, USA, June 2019.

[63] Z. Wu, Li Su, and Q. Huang, "Stacked cross refinement network for edge-aware salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7264–7273, Seoul, South Korea, November 2019.

[64] Z. Wu, Li Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3907–3916, Long Beach, CA, USA, June 2019.

[65] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, https://arxiv.org/abs/1409.1556.