



Research Article

Improved U-Net-Like Network for Visual Saliency Detection Based on Pyramid Feature Attention

Xiaoran Gong,¹ Letu Qingge,² Qing Liu ,^{3,4} and Pei Yang ¹

¹Department of Computer Technology and Application, Qinghai University, Xining 810016, China

²Department of Computer Science, North Carolina A&T State University, Greensboro, NC 27411, USA

³School of Electronic and Information Engineering, West Anhui University, Lu'an 237012, China

⁴School of Mathematics and Big Data, Anhui University of Science and Technology, Huainan 232001, China

Correspondence should be addressed to Pei Yang; yangpeinmgdx@sina.com

Received 17 May 2022; Accepted 29 July 2022; Published 22 August 2022

Academic Editor: Anandakumar H

Copyright © 2022 Xiaoran Gong et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

As a widely used technology, visual saliency detection has attracted a lot of attention in the past decades. Although a large number of methods, especially fully convolutional neural network- (FCN-) based approaches, have been proposed and achieved remarkable performance, it is still of great value to extend representative architecture to visual saliency detection task. In this paper, we propose an improved U-Net-like network, pyramid feature attention-based U-Net-like (PFAU-Net) for visual saliency detection problem. The main improvements of the proposed model include that in order to enable the network to extract features with more representation ability, we introduce a context-aware feature extraction (CFE) module and a channel attention module into the U-shaped backbone to obtain valuable multiscale features, and a feature pyramid path is also utilized in the decoder part of the network to take advantages of multilevel information. Moreover, we construct the loss function using three terms including pixel-level cross-entropy, image-level intersection over union (IoU), and a structural similarity term, which aim to make the model learn more saliency related knowledge. To verify the effectiveness of the proposed model, we conduct extensive experiments on six widely used public datasets, and the experimental results indicate that (1) our improved model can significantly improve the performance of the backbone network on all test datasets, and (2) our proposed model can outperform comparison FCN-based networks and nonneural network approaches. Both objective and qualitative evaluations verify the effectiveness of our proposed model.

1. Introduction

As an initial step of many computer vision tasks, visual saliency detection is widely used in a vast range of computer vision application fields such as object detection [1], visual tracking [2], image retrieval [3], and image semantic segmentation. Visual saliency detection, which is inspired by the ability of the human vision system (HVS) to quickly focus on impressive regions, is aimed at locating important areas of natural images. It has attracted extensive attention from researchers, and much progress has been made in the past decades.

To solve the saliency segmentation problem, many saliency algorithms have been proposed to distinguish

salient objects from irrelevant backgrounds [4–19]. Early saliency detection significantly relies on artificially designed low-level features and various prior knowledge to determine saliency. These methods usually focus on low-level visual features, and it is difficult to obtain satisfactory results in images with complex scenes. With the development of deep learning technique, convolution neural network (CNN) attracts many researchers' attention and is widely used in various vision-related fields including saliency detection. Compared to methods based on artificial features and prior knowledge, CNN-based frameworks have made significant progress in exploiting high-level semantic features [7–9]. As representative ability of features significantly affects the performance of algorithms, it is worth exploring models that

take advantages of multilevel features and other helpful cues like contextual information for saliency detection task. In addition, although many end-to-end models have been proposed with the advent of fully convolutional neural network (FCN) [20], it is still valuable to introduce and develop typical FCN models, e.g., U-Net [21], for other tasks into saliency detection. In this paper, we concentrate our efforts on expanding and improving the U-Net-like FCN network to the visual saliency detection task. We try to improve the network's representation ability by considering the different characteristics of the encoder and decoder, as well as high-level features and low-level features, and we propose an improved U-Net-like based on a pyramid feature attention strategy with the U-shaped encoder-decoder architecture in [22] as a backbone. More specifically, for the deep layers of the encoder, we introduce a context-aware pyramid feature module to obtain multiscale and multireceptive field high-level features, and then, the channel attention module is adopted to integrate different scales and receptive fields by assigning larger weight to channels conducive to saliency detection. In addition, in order to effectively extract high-level semantic information related to salient objects, we further construct a feature pyramid fusion path for the decoder to extract multilevel semantic information related to saliency targets. In summary, the main contributions of this paper are as follows:

- (1) In this study, we are committed to exploring the use of U-Net-like architecture in the visual saliency detection task and propose an improved U-Net-like network by considering the difference between high-level and low-level features, as well as the characteristics of encoder and decoder
- (2) Context-aware pyramid feature module and channel attention module are employed to help the encoder to obtain contextual information, and a feature pyramid path is added to the decoder to extract high-level semantic information by aggregating the multi-level outputs of the decoder
- (3) Experiments are conducted on six challenging datasets to test our improved U-Net-like model, and the experimental results indicate the effectiveness of the proposed network on the visual saliency detection task

The rest of the paper is organized as follows: Section 2 reviews the related works, followed by introducing the proposed method in Section 3, and Section 4 presents the experimental results and discussion, and we finally give the conclusions in Section 5.

2. Related Works

The existing saliency detection methods can be roughly divided into conventional methods and deep neural network-based methods. Conventional saliency detection approaches are mainly nonneural network methods [15, 16, 23], and they usually use low-level handcrafted features

to estimate the saliency maps of images. For example, Cheng et al. [15] proposed a global contrast method using the histogram of pixel color to construct region contrast. In literature [24], local texture patterns, color distribution, and contour information were combined to encode superpixels for region contrast computing. In addition to handcrafted features, heuristic priors are also often employed for saliency detection. Typical priors include contrast prior [25, 26], center prior [25], and boundary prior (background prior) [10], and these priors are used to identify initial foreground or background candidate regions. Although conventional methods have achieved good performance, handcrafted features and heuristic priors are difficult to acquire high-level global semantic knowledge about objects.

In recent years, deep convolutional neural networks (CNN) have achieved remarkable performance on various vision tasks. Different from many traditional algorithms that rely on low-level handcrafted features, CNN can effectively learn high-level semantic features with stronger representation ability from raw data automatically. In terms of visual saliency detection tasks, early CNN-based methods used CNN to learn high-level semantic features [27–29] and achieved superior performance compared with traditional algorithms using handcrafted features. However, in all these methods, CNN only plays the role of feature extractor, which extracts features from patches of the processed image for further classification or regression. It means that, on one hand, saliency maps generated by these methods are patch-level rather than pixel-level which may increase the overhead of the algorithms and make the boundary of the saliency map rough. On the other hand, CNN used in these methods need to be pretrained first, which is usually completed using datasets for visual recognition tasks. The advent of fully convolutional neural network (FCN) [20] provides a new way for end-to-end pixel-level saliency detection. FCN is first proposed for semantic segmentation, and it integrates feature extraction and pixel label prediction together using one network consisting of convolutional layers and deconvolutional layers. After that, a large number of FCN-based saliency detection models have been proposed, such as recurrent fully convolutional networks (RFCN) [11], deep contrast learning (DCL) [12], and deep uncertain convolutional features (UCF) [13], and they have significantly improved the performance of visual saliency detection algorithms. Although the existing FCN-based saliency detection methods have made great progress, it is still valuable to explore visual saliency detection using some FCN-based models designed for other tasks. One representative model is U-Net [21], which is a well-known FCN-based segmentation network for medical images. As an FCN-based model, U-Net has strong feature representation ability and can gradually supplement the feature information from the encoder to the decoder. Given the impressive performance of U-Net in many computer vision tasks, in this paper, we try to explore the U-Net-like architecture for the task of visual saliency detection.

The key of the FCN-based visual saliency detection algorithm is to obtain strong feature representation, and recent works have proposed different strategies to improve the

feature learning or representation ability of the networks. Zhang et al. [30] proposed to aggregate multilevel convolutional features and achieved a more accurate salient object labelling. Zhang et al. [31] designed a bidirectional message-passing structure to pass messages between multilevel features. In literature [32], the authors proposed the aggregate interaction modules to integrate the features from adjacent levels and the self-interaction modules, which are embedded in each decoder, to obtain more efficient multiscale features. Wei et al. [33] proposed F³Net, which mainly consists of cross feature module (CFM) and cascaded feedback decoder (CFD), to consider differences between features generated by different convolutional layers ignored by most feature fusion strategies. In addition, some studies designed different loss functions to constrain the network to learn features corresponding to specific targets (e.g., the boundary or position of salient objects) [22, 33]. However, most of these methods only focus on combining rich feature information for better feature representation and ignore the difference between encoder and decoder. While the core function of the encoder is to extract low-level informative features such as object edges and textures, which contain detailed location information of objects, the decoder plays an important role in extracting semantic information related to object categories. It is worth exploring integrating feature maps of different decoders to improve the ability of feature representation.

Recently, attention mechanisms have been successfully used for various vision tasks including visual saliency detection due to its strong feature selection ability. Zhang et al. [31] proposed a progressive attention-guided network that sequentially generates attention features for saliency detection through the channel and spatial attention mechanisms. Zhao and Wu [34] employed spatial attention (SA) and channel-wise attention (CA) for low-level feature maps and context-aware pyramid feature maps, respectively, to help the network pay more attention to features suitable for the current sample. However, these existing methods commonly deal with multilevel features indiscriminately while fusing convolutional features. Although some methods adopt certain strategies as gate function [31] and progressive attention mechanism [35], they only select features in some certain directions and ignore the difference between high-level features and low-level features. As the saliency feature maps corresponding to low-level features usually tend to contain noise and the saliency maps corresponding to the high-level features are commonly insufficiently detailed, it is necessary to deal with the low-level features and high-level features differently to adapt their characteristics, so as to get a better saliency map.

3. The Proposed Method

In this paper, we propose an improved U-Net-like FCN network, pyramid feature attention-based U-Net-like (PFAU-Net), for the visual saliency detection task. Figure 1 shows the architecture of our proposed network, and next, we will introduce the proposed PFAU-Net in detail from the over-

view, pyramid feature module, channel attention module, feature pyramid path, and loss function.

3.1. Overview of the Proposed PFAU-Net. As shown in Figure 1, the backbone network of the proposed PFAU-Net is U-shaped with encoder on the left and decoder on the right proposed in literature [22]. Before introducing the details of our improvements, we first present the architecture of the backbone network. The backbone is composed of two symmetrical parts, i.e., the encoder and the decoder. The encoder part contains a convolutional layer as input and is followed by six stages, which consist of basic residual blocks. The input convolutional layer contains $64 \times 3 \times 3$ filters with a stride of 1. The following four stages of the backbone are directly adopted from the ResNet-34, and two extra stages are composed of a nonoverlapping max-pooling layer with size 2 and three basic residual blocks with $512 \times 3 \times 3$ filters. In addition, there is a bridge stage consisting of three convolutional layers, each of which is with 512 dilated 3×3 filters [36] and followed by a batch normalization [37] and ReLU activation function, between the encoder and decoder to help capturing global information. The decoder part of the backbone is almost symmetric with the encoder except that upsampling instead of max-pooling is used for adjacent stages of decoder.

In order to make the model learn more useful features, in our proposed PFAU-Net, several improvements have been made to the backbone network. Firstly, one more side output is added to each stage of the decoder of the backbone. Secondly, in order to capture more beneficial features, a bridge is added between the encoder and decoder. For the first two stages, the output of each encoder stage and the input of the decoder stage are concatenated, and an extra *context-aware pyramid feature extraction* module and channel attention module, which aim to obtain multiscale and multireceptive field features, respectively, are added before concatenating for the other stages. Moreover, a feature pyramid path is added for the decoder to extract multilevel semantic information related to saliency targets, and more details about the feature pyramid path can be found in Feature Pyramid Path.

3.2. Context-Aware Pyramid Feature Extraction Module. Visual context is of great help to better represent visual content. However, most existing saliency detection models, which extract visual features by directly stacking multiple convolutional and pooling layers, rarely consider the visual context. As the scale, shape, location, etc., of salient objects in images vary greatly, it is necessary to use more representative features for the visual saliency detection task. Scale-invariant feature transformation, which is abbreviated as SIFT, is a well-known image feature descriptor proposed by Lowe [38]. It proposes the Laplacian algorithm for Gaussian representation, which fuses scale-space representation and pyramidal multiresolution representation [38]. The scale-space representation is obtained by convolution between multiple Gaussian kernel functions and images with the same resolution, while the pyramid multiresolution representation is the results of downsampling feature maps of

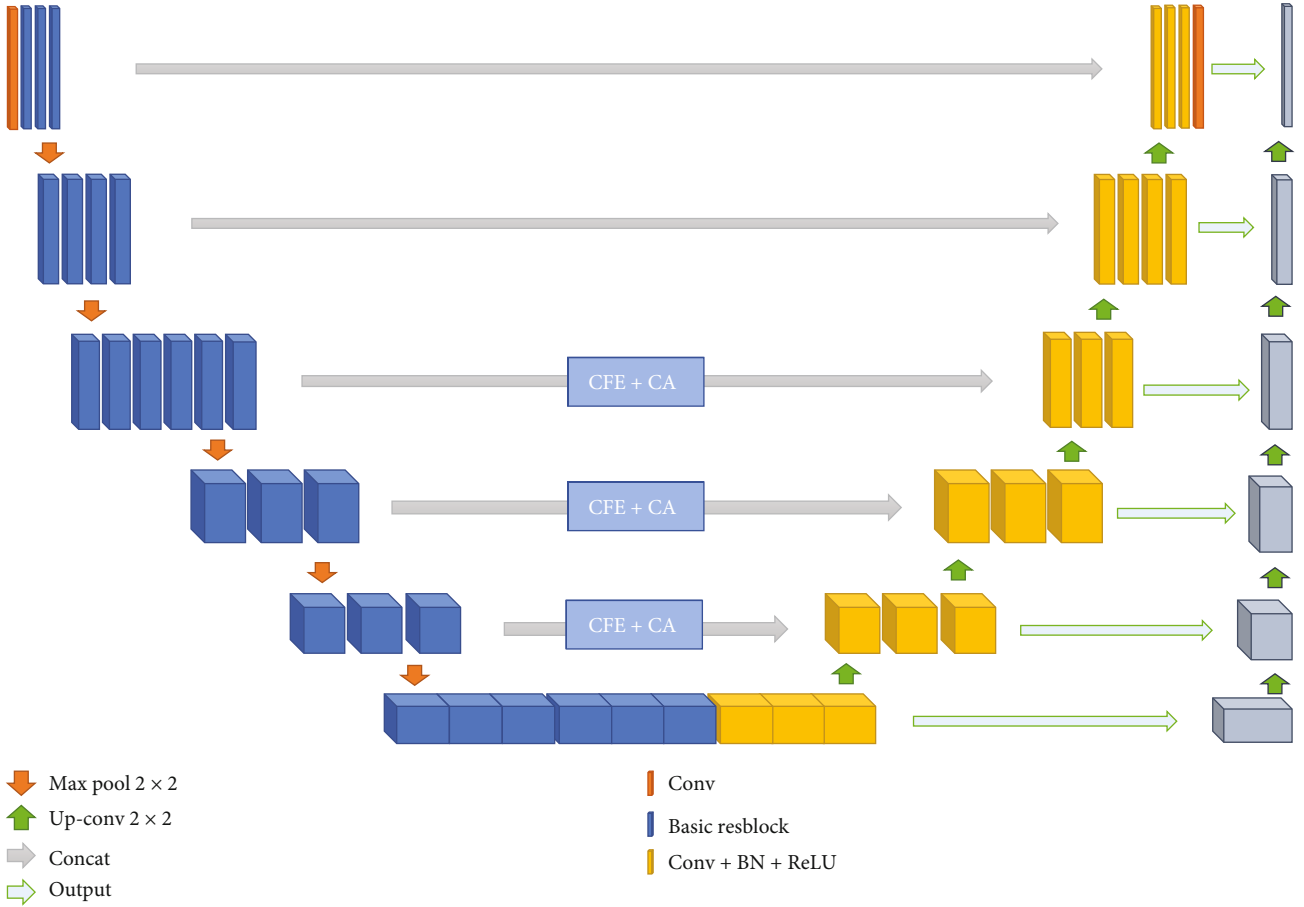


FIGURE 1: Architecture overview of our proposed pyramid feature attention-based U-Net-like (PFAU-Net) model. CFE and CA stand for context-aware feature extraction and channel attention, respectively. Compared with the U-Net-like backbone used in [22], two main improvements are made, including (1) introducing context-aware pyramid feature module into deep layers of the encoder to pass multiscale and multireceptive field high-level features to the decoder through shortcut and (2) adding a feature pyramid fusion path for the decoder to extract multilevel semantic information related to saliency targets.

different resolutions. The advantages of SIFT are that it is invariable in the scale and rotation in the image changes and robust to illumination and image deformation. Inspired by SIFT, we introduce a context-aware pyramid feature extraction (CPFE) module, which use dilated convolution to generate feature maps with the same scale but different receptive field, to extract multiscale features for stage 3, stage 4, and stage 5 of the U-Net-like backbone.

Specifically, the CPFE module is shown in Figure 2. Please note that the CPFE module uses the output of stage 3 to stage 5 as input. To make the final extracted high-level features contain more context information, we employ dilated convolutions with different dilation rates. In our implementation, we use one 1×1 convolution filter and three 3×3 dilated filters with 3, 5, and 7 as their dilation rates to assemble the context-aware feature extractor, which is abbreviated as CFE in Figure 2. We then directly concatenate the feature maps generated by the four convolution operations, and the feature maps of stage 3, stage 4, and stage 5 of the U-Net-like backbone construct the pyramid features. The stacked feature maps of each stage will be fur-

ther processed by a channel attention module, which will be introduced later, and passed to the corresponding stage of the decoder.

3.3. Channel Attention Module. As introduced in the last subsection, we utilize CPFE module to extract multiscale and multireceptive field high-level features. However, the contribution of different feature map to the visual saliency detection task is not exactly the same, which implies that the results may be disturbed if we treat all feature maps indiscriminately. Therefore, it is of great significance to filter out feature maps those contribute little to the task and further emphasize feature maps with strong correlation with the task. For this purpose, in this subsection, we introduce the attention mechanism to help the model focus on those promising feature maps. Specifically, we use a channel attention (CA) module [39] for high-level features according to the characteristics of feature maps.

Figure 3 shows the architecture of the used channel attention module. It is simple and effective implementation of channel attention proposed by Hu et al. [39]. From

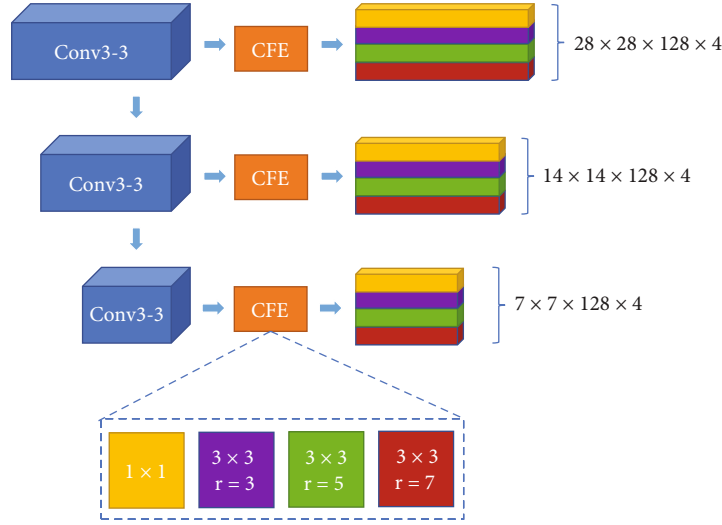


FIGURE 2: Detailed structure of the context-aware pyramid feature extraction module. The context-aware feature extraction module takes the high-level features output by the encoder of U-Net-like backbone as input and is composed of three convolutional layers with 3×3 dilated filters with different dilated rates and one 1×1 convolutional layer. CFE in this figure stands for context-aware feature extraction module.

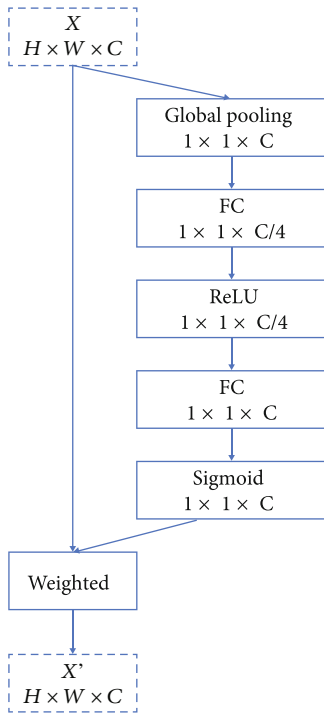


FIGURE 3: Scheme of the channel attention module used in this work [39]. X and X' represent features before and after weighting, respectively.

Figure 3, we can see that the weight generation part of the CA module is composed of a global pooling, two fully connected layers with ReLU and Sigmoid as activation function, respectively. Let $X = [x_1, x_2, \dots, x_C] \in \mathbb{R}^{H \times W \times C}$ represent the C channel input feature maps with size $H \times W$ and $avg(\cdot) : \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}^{1 \times 1 \times C}$ be a global average pooling function.

Then, we get $v \in \mathbb{R}^{1 \times 1 \times C}$ after the global pooling layer as

$$v = \text{average}(X). \quad (1)$$

Next, v is further processed by two fully connected layers $f_{c_1}(\cdot)$ and $f_{c_2}(\cdot)$ in turn, and then, we get the channel-wise weights $W_C \in \mathbb{R}^{1 \times 1 \times C}$ as

$$W_C = \sigma(f_{c_2}(\delta(f_{c_1}(v, W_1, b_1)), W_2, b_2)) = \sigma(W_2 \delta(W_1 v + b_1) + b_2), \quad (2)$$

$$\delta(x) = \max(0, x) = \begin{cases} x & x \geq 0, \\ 0 & x < 0, \end{cases} \quad (3)$$

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \quad (4)$$

where $\delta(x)$ and $\sigma(x)$ are the ReLU and Sigmoid activation functions, respectively, and W_1, b_1 and W_2, b_2 are weights and bias for $f_{c_1}(\cdot)$ and $f_{c_2}(\cdot)$, respectively. After we get the channel-wise weights W_C , we can input feature maps that can be rescaled $X' = [x'_1, x'_2, \dots, x'_C] \in \mathbb{R}^{H \times W \times C}$ as

$$X' = W_C \otimes X, \quad (5)$$

where \otimes is element-wise multiplication. Note that the channel-wise weights are input specific, which makes the CA module capable of introducing dynamics conditioned on the input and helps to boost feature maps discriminability.

3.4. Feature Pyramid Path. As semantic features corresponding to different levels of the decoder may represent different aspects of the saliency objects, considering more high-level features may promote the saliency detection results. For this

TABLE 1: Description of the six used benchmark datasets.

Dataset	Year	Publication	Image amount	Object area (%)	Annotation
SOD	2010	CVPR-W	300	27.99 ± 19.36	Pixel-wise object level
ECSSD	2015	CVPR-W	1,000	23.51 ± 14.02	Pixel-wise object level
DUT-OMRON	2013	CVPR	5,168	14.85 ± 12.15	Pixel-wise object level
PASCAL-S	2014	CVPR	850	24.23 ± 16.70	Pixel-wise object level
HKU-IS	2015	CVPR	4,447	19.13 ± 10.90	Pixel-wise object level
DUTS	2017	CVPR	15,572	23.17 ± 15.52	Pixel-wise object level

TABLE 2: Learning parameters used in the implementation.

Parameter	Value
Learning rate (lr)	0.001
Betas	(0.9, 0.999)
Eps	1e-8
Weight decay	0
Batch size	8

purpose, we add a feature pyramid path to the decoder of the backbone, which fuses output of different decoder layers as shown in Figure 1. As feature maps of the same resolution contain information with different granularity, we can get multilevel feature representation by fusing these features. We can fulfil feature fusion for adjacent feature maps in the feature pyramid path by two ways, which are “concatenate” and “add” fusion. In this study, we use an “add” connection. Suppose $X = [x_1, x_2, \dots, x_C]$ and $Y = [y_1, y_2, \dots, y_C]$ are two group of feature maps to be fused, K_i is i -th kernel for convolution, and then, the “add” fusion can be defined as

$$Z_{\text{add}} = \sum_{i=1}^c (X_i + Y_i) * K_i = \sum_{i=1}^c X_i * K_i + \sum_{i=1}^c Y_i * K_i. \quad (6)$$

3.5. Loss Function. Loss function plays a significantly important role in optimizing a machine learning model. Cross-entropy loss is one of the most used loss for classification problem, and as a binary classification problem, we can use the binary cross-entropy (BCE) loss [40] between ground truth and the predicted saliency map in pixel level. The BCE loss function is defined as follows:

$$\mathcal{L}_{\text{BCE}} = - \sum_{(r,c)} [G(r,c) \log(S(r,c)) + (1 - G(r,c)) \log(1 - S(r,c))], \quad (7)$$

where $G(r,c) \in \{0, 1\}$ is the ground truth label of the pixel at (r,c) and $S(r,c)$ is the predicted probability of the pixel at (r,c) be saliency.

Although pixel-level loss is suitable for computer processing, high-level cues (e.g., texture and shape) are more important for human beings to understand an image. In addition to BCE loss, we employed two other losses, which are SSIM loss and IoU loss. The SSIM loss function, which

can capture structural information in images, was originally proposed for image quality assessment [41]. In this paper, the SSIM loss function is used to evaluate the structural similarity of the predicted saliency map and the ground truth. Let $X = \{x_j : j = 1, \dots, N^2\}$ and $Y = \{y_i : i = 1, \dots, N^2\}$ be the two pixel blocks with size $N \times N$ cropped from the predicted probability map and the ground truth, respectively, μ_x, σ_x and μ_y, σ_y be the mean value and variance of X and Y , σ_{xy} be denoted as the covariance of X and Y , and C_1 and C_2 be the two constant. Then, the SSIM loss function of X and Y is defined as

$$\mathcal{L}_{\text{SSIM}} = 1 - \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \quad (8)$$

IoU loss function, which is an image-level metric, was originally proposed to measure the similarity of two sets [42] and later used as a standard evaluation measure for object detection and segmentation. In this study, we use IoU loss to evaluate the predicted saliency result in image-level, and it is defined as follows.

$$\mathcal{L}_{\text{IoU}} = 1 - \frac{\sum_{r=1}^H \sum_{c=1}^W S(r,c)G(r,c)}{\sum_{r=1}^H \sum_{c=1}^W [S(r,c) + G(r,c) - S(r,c)G(r,c)]}, \quad (9)$$

where $G(r,c)$ and $S(r,c)$ are the ground truth label and predicted probability of saliency of pixel at (r,c) .

Finally, we use the weighted sum of the above losses as the total loss l for a saliency map, and it can be defined using

$$l = \lambda_1 \mathcal{L}_{\text{BCE}} + \lambda_2 \mathcal{L}_{\text{SSIM}} + \lambda_3 \mathcal{L}_{\text{IoU}}, \quad (10)$$

where λ_1, λ_2 , and λ_3 are weights for BCE loss, SSIM loss, and IoU loss, and throughout our implementation, we set $\lambda_1 = \lambda_2 = \lambda_3 = 1$.

To achieve a better training effect, we use the sum of the losses of all side outputs as the final training loss \mathcal{L} , which is defined as

$$\mathcal{L} = \sum_{k=1}^K \alpha_k l^{(k)}, \quad (11)$$

where $l^{(k)}$ is the loss for the k -th side output saliency map of the decoder, K represents the total number of output losses,

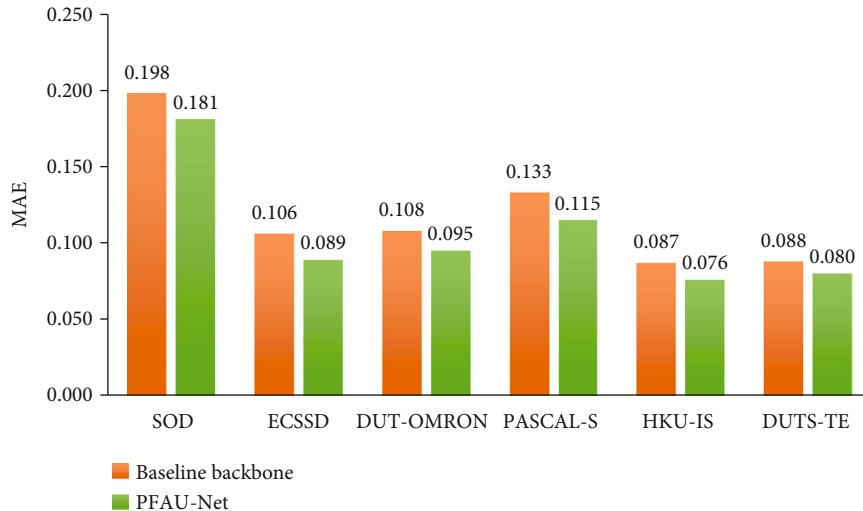


FIGURE 4: MAE of the baseline backbone and PFAU-Net on six test datasets; a smaller MAE value indicates a better performance.

TABLE 3: MAE values for the proposed model and three comparison FCN-based models on five test datasets. The results for RFCN, DCL, and UCF are from literature [35]. A smaller MAE value indicates a better performance, and the best result is highlighted in bold.

	ECSSD	DUT-OMRON	PASCAL-S	HKU-IS	DUTS-TE	Average
RFCN	0.109	0.111	0.133	0.089	0.090	0.106
DCL	0.151	0.157	0.181	0.136	0.149	0.155
UCF	0.080	0.132	0.127	0.074	0.117	0.106
Ours	0.089	0.095	0.115	0.076	0.080	0.091

and α_k is the weight of each loss. The image saliency detection model used in this paper is supervised by six side outputs.

4. Results and Discussion

4.1. Datasets and Evaluation Metric. To verify the performance of the proposed model, six widely used public benchmark datasets, including SOD [43], ECSSD [17], DUT-OMRON [16], PASCAL-S [44], HKU-IS [27], and DUTS [45], are selected for performance evaluation. Each dataset has its own characteristics; for example, SOD is an earlier one, and many images in it contain multiple salient objects similar to the background; HKU-IS consists of 4447 images with discontinuous and different spatial distribution salient objects. It should be noted that DUTS, which is currently the largest dataset, is composed of two subsets DUTS-TR and DUTS-TE with 10553 and 5019 images, respectively. More details of the six benchmark datasets are shown in Table 1. As the six used public datasets provide varied test images and pixel-level ground true, they can comprehensively evaluate the performance of saliency detection algorithms.

In order to objectively evaluate the performance of the proposed model, mean absolute error (MAE) is adopted as

the evaluation metrics. MAE is the average pixel-wise absolute difference between the predicted visual saliency map and the ground truth, and a smaller MAE value indicates a better result. Given a predicted saliency map S and ground truth G , MAE can be defined as follows:

$$\text{MAE}(S, G) = \frac{1}{H \times W} \sum_{r=1}^H \sum_{c=1}^W |S(r, c) - G(r, c)|, \quad (12)$$

where H and W are the height and width of the image, respectively, $S(r, c)$ represents the saliency value of pixel at (r, c) in the predicted visual saliency map, and $G(r, c)$ is the ground truth. For a test set, $\mathcal{D} = \{(S_i, G_i) | i = 1, \dots, N\}$ contains N test images, and its MAE is the average MAE of all test images defined as

$$\text{MAE}(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(S, G) \in \mathcal{D}} \text{MAE}(S, G). \quad (13)$$

4.2. Implementation Details. The proposed network is implemented based on the Pytorch framework. We train and test the proposed model using a Tesla P100 GPU with 16 GB video memory. The DUTS-TR dataset, which is a subset of DUTS [45, 46] and contains 10553 images, is used as the training set throughout the experiments. Please note that we augment the training set by horizontal flipping each image in DUTS-TR, which doubles the amount of training images to 21106. In addition, before feeding into the network, training images are resized to 256×256 and then cropped to 224×224 during training. To optimize the model, the Adam optimizer with default hyperparameter values is adopted. Other optimization parameters such as learning rate (lr) and betas are listed in Table 2. We train the network for about 400 K iterations to make the loss converge. When we test an image using the trained network, we first resize it to 256×256 before input into the network, and the predicted saliency map is resized back to its original size using bilinear interpolation.

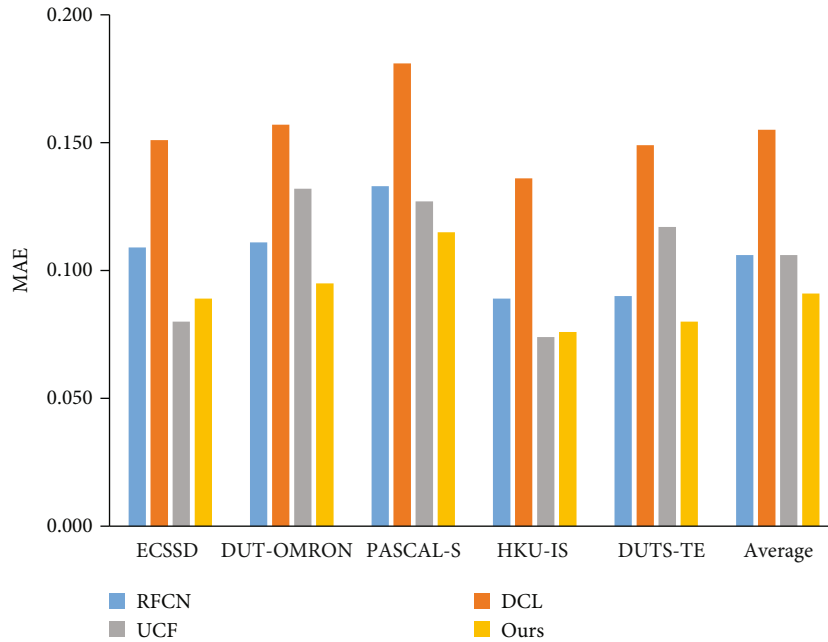


FIGURE 5: MAE of RFCN, DCL, UCF, and the proposed model on five test datasets and the average MAE of each method over the five datasets. A smaller MAE value indicates a better performance.

TABLE 4: MAE values for each method. A smaller MAE value indicates a better performance, and the best result is highlighted in bold.

	ECSSD	DUT-OMRON	PASCAL-S	SOD	Average
GS	0.206	0.173	0.221	0.251	0.213
BD	0.171	0.144	0.199	0.230	0.186
RC	0.301	0.290	0.312	0.326	0.307
MR	0.189	0.187	0.221	0.259	0.214
HS	0.228	0.229	0.262	0.283	0.251
WMR	0.191	0.201	0.234	0.265	0.223
MC	0.169	0.142	0.195	0.230	0.184
Ours	0.089	0.095	0.115	0.181	0.120

4.3. Experimental Results and Discussion. In this section, we evaluate the performance of the proposed model both quantitatively and qualitatively. Firstly, we conduct experiments to compare the performance of the baseline U-Net-like backbone network [22] and the proposed PFAU-Net. MAE values of the baseline backbone network and the proposed PFAU-Net are shown in Figure 4. The comparison results indicate that PFAU-Net improves the performance of the baseline backbone network on all the six test datasets. PFAU-Net reduced the MAE of the baseline by an average of 11.98% on the six datasets, with the largest reduction of 16.04% on ECSSD and the least on SOD (8.59%).

We further compare the results of the proposed model with some FCN-based networks including RFCN [11], DCL [12], and UCF [13]. MAE values for each comparison model are reported in Table 3. From the results, we can see that our proposed model outperforms the comparison methods on DUT-OMRON, PASCAL-S, and DUTS-TE;

the MAE values of our model are much smaller than the second smallest MAEs. On datasets ECSSD and HKU-IS, our model can obtain competitive results that the MAE values on both datasets (0.089 on ECSSD and 0.076 on HKU-IS) are only larger than those of UCF (0.080 on ECSSD and 0.074 on HKU-IS). In addition, we compute the average MAE of each model on the five test datasets, and our model achieves the best average MAE value (0.091). The results are also presented in Figure 5, from which we can obviously find that the proposed method outperforms the comparison algorithms according to average MAE value.

To further evaluate the model, we compared our proposed model with several nonneural network methods including GS [10], BD [14], RC [15], MR [16], HS [17], WMR [18], and MC [19]. MAE values of all the comparison methods are the results reported in literature [19], and Table 4 presents the results of the proposed model and all comparison methods. From Table 4, we can see that our proposed FCN-based model outperforms all the comparison methods on the four test datasets, and the average MAE of our model on the four test datasets is 0.120, which is significantly better than that of the best nonneural network method MC (0.184). The results not only verify the effectiveness of our proposed model but also further show the great advantages of neural network-based method over traditional approaches in visual saliency detection task.

In addition to objective evaluation, we also present some predicted saliency maps of the backbone network and our model in Figure 6 for qualitative evaluation. Images in the first and second row of Figure 6 are input original images and the corresponding ground truth saliency maps, respectively. The third and fourth rows are predicted saliency maps of the U-Net-like backbone network and our model. The results indicate that our model generates saliency maps more

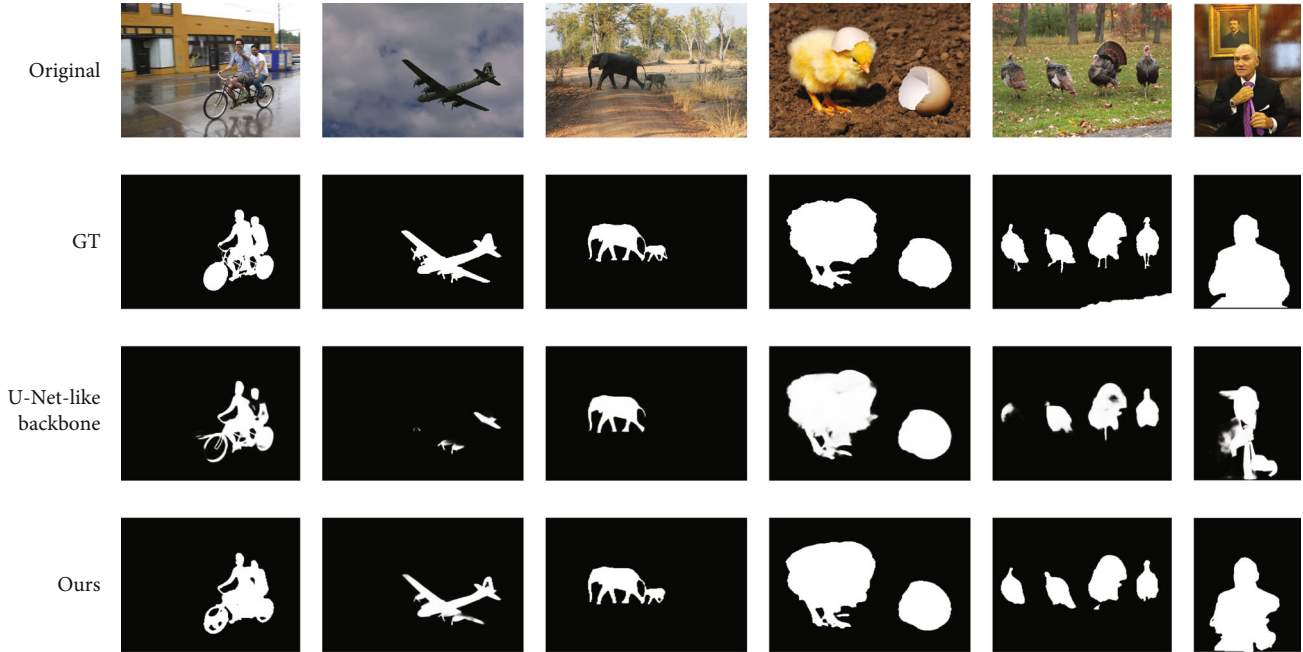


FIGURE 6: Saliency maps of our proposed PFAU-Net model and the U-Net-like backbone network. GT represents the ground truth.

accurately for different challenging scenes. For example, the third and fifth rows contain more than one salient objects, and while our model predicts all the salient objects; the baseline method misses part of the salient objects (third and fifth columns of the saliency maps results of the U-Net-like backbone). Moreover, for the second and the sixth test images, the salient objects detected by the U-Net-like backbone are seriously incomplete as shown in third row of Figure 6. However, our model can produce more accurate results compared to the backbone network. Meanwhile, we should also pay attention to the limitations of the proposed method in some special cases. Taking the fifth and sixth saliency maps of our model as an example, the legs of the animals are not detected as they are thin and long, and an incomplete body region is detected for the sixth image due to the color similarity to background.

From the above objective and qualitative results, we can find that the improved U-Net-like network can produce competitive results on saliency detection problem. The results also indicate the importance of improving feature representation as the proposed model is mainly focusing on extracting more representative features. In addition, the comparison results with nonneural methods further verify the great advantages of neural network models in saliency detection task. Although the proposed method has achieved competitive results, the limitation of the proposed model is also obvious that the boundary of the salient objects in the generated saliency map is not accurate and fine enough, and more efforts are needed in the future research.

5. Conclusion

In this paper, we propose an improved U-Net-like model PFAU-Net for visual saliency detection task. A U-shaped

encoder-decoder network is used as backbone, and in order to make the network be able to capture more useful features, a CFE module followed by a channel attention module is added to the backbone to capture multiscale features. In addition, a feature pyramid path is introduced to the decoder part to take advantages of multilevel information. To evaluate the performance of the proposed model, we compare our method with some FCN-based and nonneural methods using six widely used public datasets, and the subjective and objective results show that our proposed model has achieved competitive results, which verifies the effectiveness of the proposed method.

Data Availability

All datasets supporting this study are publicly available by applying to the datasets owner.

Conflicts of Interest

The authors declare that they have no conflicts of interest in this work.

Acknowledgments

This work was partly supported by the National Natural Science Foundation of China (Nos. 61866031 and 62102002), Science Technology Foundation for Middle-Aged and Young Scientist of Qinghai University (No. 2018-QGY-6), and Natural Science Foundation of Anhui Province (No. 2008085QF291).

References

- [1] Y. Ding, J. Xiao, and J. Yu, "Importance filtering for image retargeting," in *IEEE conference on computer vision and pattern recognition*, pp. 89–96, Colorado Springs, CO, USA, 2011.
- [2] A. Borji, S. Frintrop, D. N. Sihite, and L. Itti, "Adaptive object tracking by learning background context," in *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pp. 23–30, Providence, RI, USA, June 2012.
- [3] Y. Gao, M. Wang, D. Tao, R. Ji, and Q. Dai, "3-D object retrieval and recognition with hypergraph analysis," *IEEE Transactions on Image Processing*, vol. 21, no. 9, pp. 4290–4303, 2012.
- [4] N. Kousik, Y. Natarajan, R. A. Raja, S. Kallam, R. Patan, and A. H. Gandomi, "Improved salient object detection using hybrid convolution recurrent neural network," *Expert Systems with Applications*, vol. 166, article 114064, 2021.
- [5] A. Siris, J. Jiao, G. K. Tam, X. Xie, and R. W. Lau, "Scene context-aware salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4156–4166, Montreal, QC, Canada, 2021.
- [6] M. Zhuge, D. P. Fan, N. Liu, D. Zhang, D. Xu, and L. Shao, "Salient object detection via integrity learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, 2022.
- [7] H. Li, J. Chen, H. Lu, and Z. Chi, "CNN for saliency detection with low-level feature integration," *Neurocomputing*, vol. 226, pp. 212–220, 2017.
- [8] C. Dai, C. Pan, and W. He, "Feature extraction and fusion network for salient object detection," *Multimedia Tools and Applications*, vol. 81, pp. 1–15, 2022.
- [9] Y. Ji, H. Zhang, Z. Zhang, and M. Liu, "CNN-based encoder-decoder networks for salient object detection: a comprehensive review and recent advances," *Information Sciences*, vol. 546, pp. 835–857, 2021.
- [10] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *European conference on computer vision*, Springer, Berlin, Heidelberg, 2012.
- [11] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *European Conference on Computer Vision*, pp. 825–841, Springer, Cham, 2016.
- [12] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 478–487, Las Vegas, NV, USA, 2016.
- [13] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *Proceedings of the IEEE International Conference on computer vision*, pp. 212–221, Venice, Italy, 2017.
- [14] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2814–2821, Columbus, OH, USA, 2014.
- [15] 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 Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2015.
- [16] C. Yang, L. Zhang, H. Lu, X. Ruan, 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, 2013.
- [17] 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, 2013.
- [18] X. Zhu, C. Tang, P. Wang et al., "Saliency detection via affinity graph learning and weighted manifold ranking," *Neurocomputing*, vol. 312, pp. 239–250, 2018.
- [19] Q. R. Zhang and Y. F. Wang, "A multi-cues based approach for visual saliency detection," *International journal of innovative computing, information and control*, vol. 17, pp. 1435–1446, 2021.
- [20] 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, 2015.
- [21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, Springer, Cham, 2015.
- [22] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: boundary-aware salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7479–7489, Long Beach, CA, USA, 2019.
- [23] M. Donoser, M. Urschler, M. Hirzer, and H. Bischof, "Saliency driven total variation segmentation," in *2009 IEEE 12th International Conference on Computer Vision*, pp. 817–824, Kyoto, Japan, 2009.
- [24] B. Nan, Z. Mu, L. Chen, and J. Cheng, "A local texture-based superpixel feature coding for saliency detection combined with global saliency," *Applied Sciences*, vol. 5, no. 4, pp. 1528–1546, 2015.
- [25] C. Yang, L. Zhang, and H. Lu, "Graph-regularized saliency detection with convex-hull-based center prior," *IEEE Signal Processing Letters*, vol. 20, no. 7, pp. 637–640, 2013.
- [26] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1915–1926, 2012.
- [27] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5455–5463, Boston, MA, 2015.
- [28] L. Wang, H. Lu, X. Ruan, 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, 2015.
- [29] 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, 2015.
- [30] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: aggregating multi-level convolutional features for salient object detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 202–211, Venice, Italy, 2017.
- [31] L. Zhang, J. Dai, H. Lu, Y. He, 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, 2018.

- [32] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9413–9422, Seattle, WA, USA, 2020.
- [33] J. Wei, S. Wang, and Q. Huang, "F³Net: fusion, feedback and focus for salient object detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 12321–12328, 2020.
- [34] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3085–3094, Long Beach, CA, USA, 2019.
- [35] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 714–722, Salt Lake City, UT, USA, 2018.
- [36] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, <http://arxiv.org/abs/1511.07122>.
- [37] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," *International conference on machine learning*, vol. 37, pp. 448–456, 2015.
- [38] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [39] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, Salt Lake City, UT, USA, 2018.
- [40] 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.
- [41] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers*, pp. 1398–1402, Pacific Grove, CA, USA, November 2003.
- [42] G. Mátyus, W. Luo, and R. Urtasun, "Extracting road topology from aerial images," in *Proceedings of the IEEE international conference on computer vision*, pp. 3438–3446, Venice, Italy, 2017.
- [43] V. Movahedi and J. H. Elder, "Design and perceptual validation of performance measures for salient object segmentation," in *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*, pp. 49–56, San Francisco, CA, USA, June 2010.
- [44] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 280–287, Columbus, OH, USA, 2014.
- [45] L. Wang, H. Lu, Y. Wang et al. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 136–145, Boston, MA, 2017.
- [46] 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, 2013.