Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 2472676, 12 pages
https://doi.org/10.1155/2021/2472676

Research Article

Multistrengthening Module-Based Salient Object Detection

Qian Zhao,[1,2] Haifeng Wang,[1,2] Junpeng Dang,[3] Songlin Li,[3] Rong Chang,[3] Yanbin Fang,[3] Zhi Zhang,[3] Jie Peng,[3] and Yang Yang[1,2]

[1]*School of Information Science and Technology, Yunnan Normal University, Kunming 650500, China*
[2]*Laboratory of Pattern Recognition and Artificial Intelligence, Yunnan Normal University, Kunming 650500, China*
[3]*Yuxi Power Supply Bureau, Yunnan Power Grid Co Ltd, Yuxi 653100, China*

Correspondence should be addressed to Junpeng Dang; 540967252@qq.com and Yang Yang; yyang_ynu@163.com

Received 21 August 2021; Accepted 4 October 2021; Published 2 November 2021

Academic Editor: Jiayi Ma

Object detection is a classical research problem in computer vision, and it is widely used in the automatic monitoring field of various production safety. However, current object detection techniques often suffer low detection accuracy when an image has a complex background. To solve this problem, this paper proposes a double U-shaped multireinforced unit structure network (DUMRN). The proposed network consists of a detection module (DM), a reinforced module (RM), and a salient loss function (SLF). Extensive experiments on five public datasets and a practical application dataset are conducted and compared against nine state-of-the-art methods. The experiment results show the superiority of our method over the state of the art.

1. Introduction

Object detection in computer vision is widely used in the field of production safety monitoring, for example, abnormal behavior detection, regional invasion detection, and dress code detection. The practical applications of object detection effectively solve many problems and defects in safety production management while reducing preventable accidents in the workplace.

In many practical production safety applications of objection detection, we found that the following problems still exist in the safety harness detection: (1) the color contrast between safety harness and work clothes is low, which makes it difficult to detect the safety harness accurately. (2) The structure of safety harness is complex, and the detection of safety harness is easily interfered by the texture of work clothes, resulting in detection difficulty, as shown in Figure 1.

In order to solve the above problems, salient object detection technology is considered as a feasible solution. Salient object detection is to imitate the mechanism of human visual attention, that is, to obtain the object of interest in the visual scene through human eyes and then transmit this object to the brain for understanding and optimization, so as to quickly obtain the desired information from the scene. Since salient object detection can ignore irrelevant information, the region of interest can be effectively segmented and applied to the subsequent detection process. Therefore, salient object detection, as an effective pretreatment technology, is widely used in many computer vision tasks.

Based on the extensive literature review, we briefly introduce the current salient target detection methods from the following four aspects.

1.1. Salient Object Detection Based on Traditional Methods. Early salient object detection methods were based on low-level features and some heuristics' prior knowledge, such as color contrast [1], background prior [2], and center prior [3]. Early methods detect salient objects by searching for pixels according to a predefined saliency measure computed based on handcrafted features [4, 5]. Borji et al. [6] provided a comprehensive survey in this field. Encouraged by the advancement on the image classification of deep CNNs [7,8], early deep salient object detection methods searched for salient objects by classifying image pixels or superpixels into salient or nonsalient classes based on the local image patches extracted from single or multiple scales [9,10].

Figure 1: Safety harness detection for production safety monitoring.

1.2. Salient Object Detection Based on Feature Enhancement. Wang et al. [11] used a weight sharing method to refine features iteratively and promoted a mutual fusion between features. Li et al. [12] proposed a novel dense attentive feature enhancement (DAFE) module for efficient feature enhancement in saliency detection. Zhang et al. (UCF) [13] developed a reformulated dropout and a hybrid upsampling module to reduce the checkerboard artifacts of deconvolution operators as well as aggregate multilevel convolutional features for saliency detection. Hu et al. [14] proposed a level set [15] function to output accurate boundaries and compact saliency. Luo et al. (NLDF+) [16] designed a network with a 4×5 grid structure to combine local and global information and used a fusing loss of cross-entropy and boundary IoU inspired by Mumford and Shah [17]. Hou et al. (DSS+) [18] proposed a holistically nested edge detector (HED) [19] by introducing short connections to its skip layers for saliency prediction. Chen et al. (RAS) [20] presented a HED by refining its side output iteratively using a reverse attention model. Zhang et al. (LFR) [21] predicted saliency with clear boundaries by proposing a sibling architecture and a structural loss function. Yao and Wang [22] proposed an enhancing region and boundary awareness network (ERBANet) equipped with attentional feature enhancement (AFE) modules to improve the detection performance.

1.3. Salient Object Detection Based on the Attention Mechanism. The gate unit is combined by two consecutive feature maps of different resolutions from the encoder to generate rich contextual information [24]. Li et al. [5] proposed an attention steered interweave fusion network (ASIF-Net) to detect salient objects, which progressively integrated cross-modal and cross-level complementarity from the RGB image and the corresponding depth map via steering of an attention mechanism. Xu et al. [25] proposed a dual pyramid network (DPNet) for salient object detection by formulating the self-attention mechanism into the subregion-based contexts. Zhou et al. [26] proposed a simple yet effective hierarchical U-shape attention network (HUAN) to learn a robust mapping function for salient object detection and formulated a novel attention mechanism to improve the well-known U-shape network. Li et al. [27] proposed a multiattention guided feature fusion network (MAF). A novel channel-wise attention block (CAB) was used which is in charge of message passing layer by layer from a global view, which utilized the semantic cues in the higher convolutional block to instruct the feature selection in the lower block. Zhang et al. (PAGRN) [28] developed a recurrent saliency detection model that transfers global information from the deep layer to shallower layers by a multipath recurrent connection. Hu et al. (RADF+) [29] recurrently concatenated multilayer deep features for saliency object detection. Wang et al. (RFCN) [30] designed a recurrent FCN for saliency detection by iteratively correcting prediction errors. Liu et al. (PiCANetR) [31] predicted the pixel-wise attention maps by a contextual attention network and then incorporated them with U-Net.

1.4. Salient Object Detection Based on Edge Optimization. To capture finer structures and more accurate boundaries, numerous refinement strategies have been proposed. Wu et al. [32] proposed a novel stacked cross refinement network (SCRN) for salient object detection which aimed to simultaneously refine multilevel features of salient object detection and edge detection by stacking a cross refinement unit (CRU). Wang et al. (SRM) [33] proposed to capture global context information with a pyramid pooling module and a multistage refinement mechanism for saliency maps' refinement. Amirul et al. [34] proposed an encoder-decoder network that utilizes a refinement unit to recurrently refine the saliency maps from low resolution to high resolution. Deng et al. (R3Net+) [23] developed a recurrent residual refinement network for saliency maps' refinement by incorporating shallow and deep layers' features alternately. Fu et al. [35] proposed an end-to-end deep-learning-based refinement model named Refinet. Intermediate saliency maps that are edge-aware were computed from segmentation-based pooling and then fed to a two-tier fully convolutional network for effective fusion and refinement.

The researchers have improved salient object detection from the above four aspects, but the following two problems still exist.

1.4.1. The Issue of Blurred Edges. The salient object detection method based on the fully convolutional neural network (FCN) can better extract multilevel features compared with previous methods. However, after continuous convolution and pooling operations, the loss of shallow fine details cannot be reconstructed by the upsampling operation, resulting in defects in the fine structure or boundary as shown in Figure 2. The saliency is defined primarily in terms of the global features of an image, rather than local or pixel-level features. In order to obtain more accurate results, salient object detection methods still need to understand the global significance of the whole image and the structural details of the object [19].

1.4.2. The Issue of Complex Background. The most salient object detection networks adopt the U-Net structure as the encoder and the decoder, and multistage features provided by U-Net are used to reconstruct high-resolution feature images. Whether the effective features of the encoder can be transmitted to the decoder is the basis of whether the decoder can output an accurate and salient object. However, most U-Net-based methods only considered the information interaction between different levels in the encoder or the decoder and directly used all-pass skip-layer structure to connect the encoder features to the decoder. In these methods, information interference often occurs between different blocks, especially when an image has a complex background.

Qin et al. [36] proposed a method that divided the task into two parts, but optimized the edges without taking into account the loss of fine structure and the interference of the complex background. Inspired by the BASNet [36] structure, we propose a double U-shaped multireinforced unit structure network (DUMRN) to solve the above two problems simultaneously. This network can achieve a fine prediction of object boundary and accurate saliency object detection under the complex background. The main contributions of this paper include the following:

(1) We propose a new detection module which includes an information processing unit, a dual-flow branch unit, and a semantic reinforcement unit. The information processing unit is used to control the amount of information flowing from each encoder block to the decoder while enhancing the effective information and suppressing the irrelevant information. The dual-flow branch unit is used to fuse the output of the information processing unit and the supplementary branch optimizing the residual information of the trunk branch. The semantic reinforcement unit makes full use of the top-level semantic information and integrates multilevel context information to obtain more accurate spatial information and fine boundary information.

(2) We propose a new reinforcement module. It includes a feature reinforcement unit and a heat map unit. The feature reinforcement unit further fuses the information in the output preliminary salient map through a U-shaped encoder-decoder structure. The heat map unit uses the improved activation function to delimit the feature map.

(3) We design a loss function for salient object detection. It combines a salient loss of binary cross-entropy (BCE), a structural similarity (SSIM), and an IoU loss and can learn from real ground information at pixel, patch, and map levels.

2. Materials and Methods

Because encoder-decoder structures can make full use of context features in salient object detection, two encoder-decoder structures are designed to form a double U-shaped network, which is divided into the detection module (DM) and reinforcement module (RM) as shown in Figure 3.

In the previous double U-shaped network [36], the first U-shaped structure is a simple encoder-decoder structure, which often cannot effectively solve the loss of semantic information and interference of redundant information. So, we add some optimization units to the first U-encoder-decoder structure which includes an information processing unit (IPU), a dual-flow branch unit (DFBU), and a semantic reinforcement unit (SRU). We input an image into the encoder-decoder, and after the information allocation of the IPU and the information supplement of the DFBU, the output of D1 and SRU is finally added to get the preliminary feature map. Experimental results of the method with a double U-shaped structure show that adding a second encoder-decoder structure can further enhance the information [36]. The second U-shaped structure has a heat map unit (HMU) and a feature reinforcement unit (FRU). In the preliminary salient map input enhancement module, the operations of two branches are carried out, respectively. Finally, the output of $S'1$ of the FRU and HMU is added to obtain the final result.

2.1. Detection Module. The detection module is mainly aimed at solving the problems of information interference under complex background and edge blurring. The detection module is a U-shaped encoder-decoder structure, which mainly contains IPU, DFBU, and SRU. IPU controls and processes the information exchange between encoders and decoders to solve the interference problem caused by the complex background. DFBU supplements the main information and solves the problem of detailed information. SRU makes better use of multilevel semantic information to solve the problem of the edge structure.

2.1.1. Information Processing Unit (IPU). Compared with previous methods, the U-Net structure can obtain both deep semantic information and shallow spatial information. However, there is interference information when the encoder and decoder exchange information, and transmitted information has many invalid information, or the interference affects the quality of transmitting information.

To solve this problem, we add an IPU between each pair of corresponding encoders and decoders to distribute the

Figure 2: Our method is compared with the GateNet method. (a) The region of interest. (b) An enlarged view of ground truth (GT). (c) The result of GateNet. (d) The result of our method.

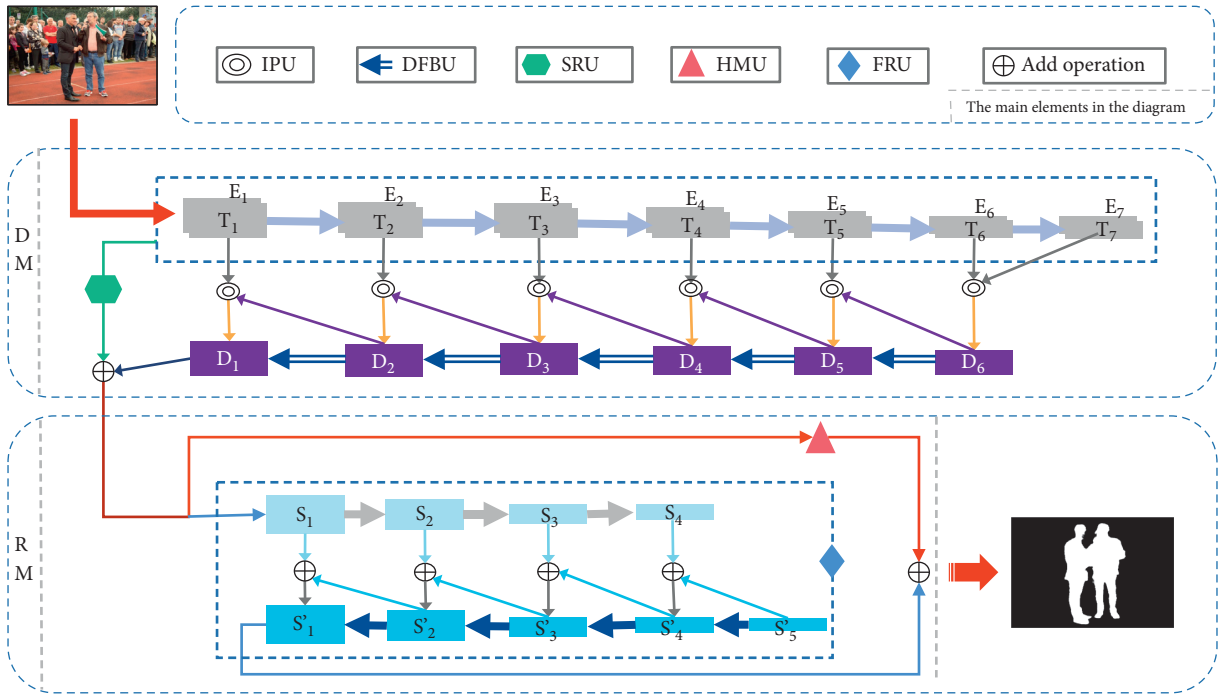


Figure 3: DUMRN architecture. The network is divided into a detection module and a reinforcement module. The detection module includes an information processing unit, a dual-flow branch unit, and a semantic reinforcement unit. The reinforcement module includes a heat map unit and a feature reinforcement unit. DM represents the detection module, RM represents the reinforcement module, E_i and S_i represent convolution layers of the encoder, D_i, and S_i' represent convolution layers of the decoder, and T_i represents transition convolution layers with the same size as E_i.

information from the encoder and then transmit it to the decoder, as shown in Figure 4. Among them, E_i represents the i-layer feature of the encoder, T_i represents the $i$layer feature parallel to the encoder, and D_{i+1} represents the decoder feature of the $i+1$ layer. When information is allocated, E_i, T_i, and D_{i+1} are input into the IPU for a series of convolution, activation, and pooling operations to obtain a specific gravity to be allocated to X_i. The operation formula is as follows:

$$X_i = P(S(\mathrm{Conv}(\mathrm{Cat}(E_i, T_i, D_{i+1})))), \quad (1)$$

where $\mathrm{Cat}(\cdot)$ is the connection operation between channels, $\mathrm{Conv}(\cdot)$ is the convolution layer, and $S(\cdot)$ is the sign function on the element.

2.1.2. Dual-Flow Branch Unit (DFBU). The DFBU structure, the trunk branch as the main fusion feature of the main trunk, is used to combine multilevel information to predict the overall information of the object, while the supplementary branch can combine more low-level information to supplement the trunk branch, aiming at optimization. The IPU processed information X_i is divided into two branches and entered into the DFBU, respectively. Part of the information enters into the main branch to obtain the convolution layer D_i. After the convolution, activation, and pooling operations, add it to the output of X_i to get the convolution level D_{i-1}. Part of the information goes to the supplementary branch, and the information at all levels is added in turn. Finally, the output information is added to the

D_1 output information of the trunk branch to obtain the DFBU output result, which is denoted as output_1, as shown in Figure 5.

2.1.3. Semantic Reinforcement Unit (SRU). Since spatial information and detailed information cannot be fully integrated in the U-Net structure, the shallow semantic information is lost step by step in the continuous process of convolution and pooling of input information. Rich semantic information and accurate detailed information play an important role in salient object detection. Due to the lack of shallow and deep features, the generated salient map cannot obtain fine boundaries in the case of satisfying the approximate salient region. Since the highest layer of the encoder has rich semantic features, we fuse the features of multiple layers in the encoder (E_2, E_3, E_4, E_5, and E_6) with E_1, respectively, to obtain a convolution layer with the same size as E_1. Finally, we add the five fused convolution layers Y_2 to Y_6 and E_1, and the output result is output_2, as shown in formula (2) and Figure 6. Finally, output_1 and output_2 are added to obtain the preliminary feature graph output by the detection module.

$$X = E_1 + \sum_{i=2}^{6} \left(E_1 + \mathrm{up}\left(E_i\right)\right). \tag{2}$$

2.2. Reinforcement Module

2.2.1. Heat Map Unit (HMU). The heat map unit mainly intensifies and weakens the features in the feature map. We introduce a nonlinear activation function, sigmoid activation function, to adjust the features in the graph. Sigmoid is also known as the logistic activation function. It compresses a real value to the range of 0 to 1. It can be applied to the output layer when our ultimate goal is to predict probability. It turns big negative numbers into zero and big positive numbers into one. We adjust this function, as shown in Figure 8, to make the graph of the function steeper and more dramatic as it approaches 0 in the x-direction. This function will strengthen salient features and suppress nonsalient features in the input image, thus forming a feature map similar to a heat map. Suppression of information on the left side of the Y-axis makes the background of the salient map cleaner.

2.2.2. Feature Reinforcement Unit (FRU). The feature reinforcement unit is the second U-shaped structure of the network: the encoder and decoder structure. HMU can cause the loss of some valid information mixed into the HMU when it restrains nonsignificant information, so the FRU mainly supplements the information in the HMU. By using the characteristics of the U-Net structure, the FRU can make better use of deep and shallow information to reinforce the features of the initial salient feature graph and finally output the results in the last convolution layer S_1' of the decoder. The output results are fused with the output results of the heat map unit, and the features of the convolution layer are further reinforced to obtain the final results, as shown in Figure 9.

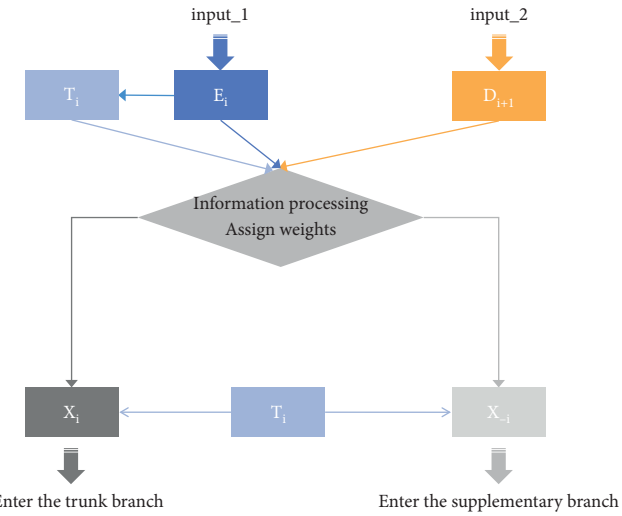


Figure 4: Information processing unit. Input_1 indicates the input of the encoder of the IPU, and input_2 indicates the input of the decoder.

2.3. Loss Function.

In order to train the salient target detection model, we design three significance loss functions according to the format of significance loss in the previous method [36], which include BCE loss, SSIM loss, and IOU loss.

$$\mathscr{L} = \sum_{i=1}^{k} \alpha^{(i)} \mathscr{L}_{\mathrm{bce}}^{(i)} + \beta^{(i)} \mathscr{L}_{\mathrm{ssim}}^{(i)} + \gamma^{(i)} \mathscr{L}_{\mathrm{iou}}^{(i)}, \tag{3}$$

where $\mathscr{L}^{(i)}$ represents the output loss at the i-side; k represents the total number of sides; $\alpha^{(i)}$, $\beta^{(i)}$, and $\gamma^{(i)}$ represent the weight of each loss of BCE, SSIM, and IOU, respectively; and $\mathscr{L}_{\mathrm{bce}}$, $\mathscr{L}_{\mathrm{ssim}}$, and $\mathscr{L}_{\mathrm{iou}}$ represent the BCE loss, SSIM loss, and IOU loss, respectively. Our model has 8 outputs, namely, $k = 8$, including 7 outputs of the detection module and 1 output of the enhancement module. $\mathscr{L}_{bce}$ is defined as

$$\begin{aligned} \mathrm{output} &= t * \log\left(\frac{1}{p} - 1\right) - \log(1 - p), \\ \mathscr{L}_{\mathrm{bce}} &= \sum_{i=1}^{k} \mathrm{output}^{(i)}, \end{aligned} \tag{4}$$

where t represents the ground truth value and p represents the predicted value.

SSIM is originally proposed for image quality evaluation. It explores structural information in an image by separating the effects of brightness on objects. The similarity measurement of SSIM can be composed of three contrast modules, respectively: brightness, contrast, and structure. $\mathscr{L}_{\mathrm{ssim}}$ is defined as

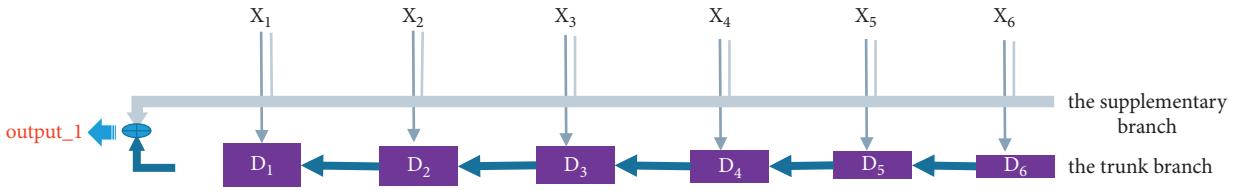


Figure 5: Dual-flow branch unit. X_i is the IPU output and DFBU input, D_i is each convolution layer of the main trunk branch, and output_1 is the output result.

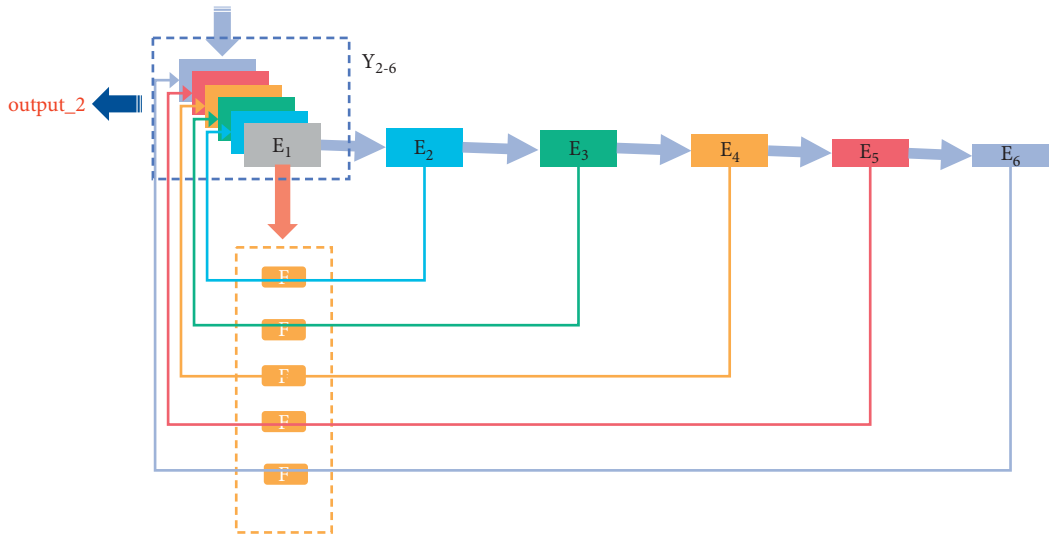


Figure 6: Semantic reinforcement unit. F is the unit that performs the fusion operation as shown in Figure 7.

$$\mathscr{L}_{ssim}(x,y) = \left(\frac{2\left(\sum_{i=1}^{N} w_i x_i\right) * \left(\sum_{i=1}^{N} w_i y_i\right) + C_1}{\left(\sum_{i=1}^{N} w_i x_i\right)^2 + \left(\sum_{i=1}^{N} w_i y_i\right)^2 + C_1}\right)^{\alpha} * \left(\frac{2\left(\left(\sum_{i=1}^{N} w_i (x_i - \mu_x)^2\right)^{1/2} * \left(\sum_{i=1}^{N} w_i \left(y_i - \mu_y\right)^2\right)^{1/2}\right) + C_2}{\left(\sum_{i=1}^{N} w_i (x_i - \mu_x)^2\right) + \left(\sum_{i=1}^{N} w_i \left(y_i - \mu_y\right)^2\right) + C_2}\right)^{\beta} * \left(\frac{\sum_{i=1}^{N} w_i (x_i - \mu_x)\left(y_i - \mu_y\right) + C_3}{\left(\sum_{i=1}^{N} w_i (x_i - \mu_x)^2\right)^{1/2} * \left(\sum_{i=1}^{N} w_i \left(y_i - \mu_y\right)^2\right)^{1/2} + C_3}\right)^{\gamma}, \tag{5}$$

where x, y represents the image feature; x_i, y_i represents the positions of the local SSIM index in the mapping; w_i represents the value of the symmetric Gaussian weighting function; and the constants C_1, C_2, and C_3 are to avoid the instability of the system caused when $\mu_x^2 + \mu_y^2$ approaches 0. α, β, and γ are greater than zero and are set to 1 in practice.

$\mathscr{L}_{\text{iou}}$ is used as a standard evaluation measure and training loss for object detection and segmentation, which can reflect the detection effect. The expression is as follows:

$$\text{IOU} = \frac{\text{Prediction} \cap \text{GroundTruth}}{\text{Prediction} \cup \text{GroundTruth}},$$
$$\mathscr{L}_{\text{iou}} = 1 - \text{IOU}. \tag{6}$$

Among them, GroundTruth is the correct result annotated artificially, while Predict represents the result predicted by the algorithm. The IOU standard is used to measure the correlation between true and predicted, and the higher the correlation, the higher the value.

We have introduced the principle and calculation process of the three loss functions. These functions represent different stages of the training process. $\mathscr{L}_{\text{bce}}$ is a pixel-level-based convergence assessment, and different weights are assigned to the foreground and background. $\mathscr{L}_{\text{ssim}}$ calculates the local neighborhood of each pixel and assigns a higher weight to the boundary to make the boundary clearer. $\mathscr{L}_{\text{iou}}$ is used to measure the correlation between real and predicted values. When combining these three kinds of loss, we use $\mathscr{L}_{\text{bce}}$ to maintain the smooth gradient of all pixels and use $\mathscr{L}_{\text{iou}}$ to pay more attention to the foreground. $\mathscr{L}_{\text{ssim}}$ is used to enhance the target boundary information in the feature map.

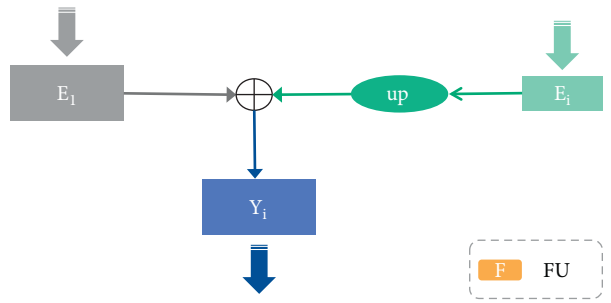


Figure 7: F: unit of the semantic reinforcement unit. The convolution layer other than E_1 is enlarged to the same size as E_1 and finally fused with E_1 to obtain Y_i.

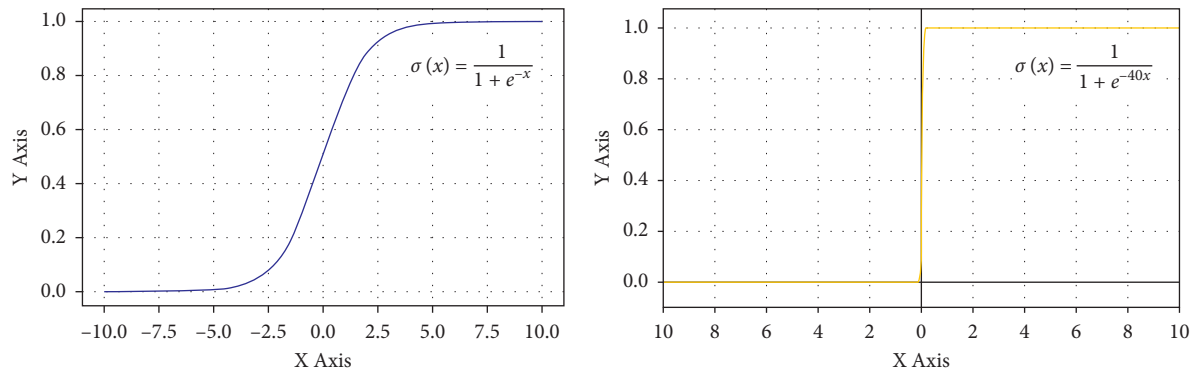


Figure 8: (a) The sigmoid activation function. (b) The improved function multiplies X by 40.

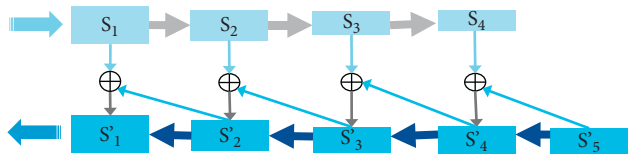


Figure 9: Feature reinforcement unit. S_i is the convolution layer of the encoder of the FRU. S_i' is the convolution layer of the decoder.

3. Experiment

3.1. Experimental Dataset. In this section, we first test the proposed method using the following five image saliency detection datasets:

1 ECSSD contains 1000 images with complex structure.

(2) DUT-OMRON contains 5168 images with complex foreground structure, each of which usually has complex background or multiple foreground objects.

(3) PASCAL-S contains 850 images of background and complex foreground objects.

(4) HKU-IS contains 4447 images with multiple foreground objects that overlap or break the image boundary.

(5) DUTS is the largest dataset for image saliency detection, which consists of two subsets: DUTS-TR and DUTS-TE. DUTS-TR contains 10553 images for training, and DUTS-TE contains 5019 images for testing.

We then apply the proposed method to the safety harness detection task for the practical application. For this practical application, we have collected 2200 images from power construction sites, 2000 for the model training and 200 for testing.

3.1.1. Implementation and Experimental Setup. We use an eight-core PC with AMD Ryzen 1800 × 3.5 GHz CPU (32 GB memory) and GTX 1080ti GPU (11 GB memory) for training and testing. We build our model on the basis of the BASNet framework. In the experiment, the proposed network is implemented on the PyTorch repository. We train the network using the DUTS-TR dataset. During training, each image is first resized to 256 × 256 (pix) and randomly cut to 224 × 224 (pix). For the optimizer, we use the Adam optimizer to train our network, and its superparameter settings

Table 1: The results of the ablation study. The experiment uses the FPN as the evaluation baseline.

Configurations	F_β	S_m	MAE
Baseline (FPN)	0.816	0.829	0.060
+IPU	0.840	0.847	0.053
+DFBU	0.870	0.869	0.045
+SRU	0.9312	0.9045	0.0424
+HMU	0.9327	0.8947	0.0454
+FRU	0.9854	0.9730	0.0068

Table 2: Quantitative evaluation on five datasets.

Method	ECSSD 1,000 images			HKU-IS 4,447 mages		
	MAE	F_β	S_m	MAE	F_β	S_m
BASNet	0.037	0.88	0.921	0.032	0.895	0.946
F3Net	0.033	0.925	0.927	0.028	0.91	0.953
GateNet	0.041	0.941	0.917	0.035	0.928	0.909
ITSD	0.04	0.939	—	0.035	0.927	—
LDF	0.034	0.93	0.925	0.027	0.914	0.954
MINet	0.036	0.943	0.947	0.03	0.932	0.955
AFNet	0.042	0.935	0.914	0.036	0.923	0.905
EGNet	0.044	0.941	0.913	0.034	0.929	0.91
PoolNet	0.038	0.945	—	0.03	0.935	—
Ours	0.028	0.985	0.993	0.012	0.968	0.987

Method	DUT-OMRON 5,168 images			DUTS-TE 5,019 images			PASCAL-S 850 images		
	MAE	F_β	S_m	MAE	F_β	S_m	MAE	F_β	S_m
BASNet	0.056	0.756	0.869	0.048	0.791	0.884	0.076	0.775	0.847
F3Net	0.053	0.766	0.87	0.035	0.84	0.902	0.062	0.84	0.859
GateNet	0.061	0.794	0.82	0.045	0.87	0.869	0.07	0.882	0.855
ITSD	0.063	0.813	—	0.042	0.877	—	—	—	—
LDF	0.051	0.773	0.873	0.034	0.855	0.91	0.06	0.848	0.865
MINet	0.057	0.794	0.864	0.039	0.877	0.912	0.065	0.882	0.898
AFNet	0.057	0.797	0.826	0.046	0.862	0.866	0.071	0.868	0.85
EGNet	0.056	0.826	0.813	0.043	0.88	0.866	0.076	0.863	0.848
PoolNet	0.053	0.833	—	0.036	0.892	—	0.065	0.88	—
Ours	0.014	0.942	0.976	0.013	0.945	0.981	0.069	0.969	0.983

are as follows: initial learning rate $lr = le - 3$, betas = $(0.9, 0.999)$, eps = $1e - 8$, and weight attenuation = 0. During testing, each input image is adjusted to 256×256 (pix) and then input into the network to obtain the saliency map. The saliency map is adjusted to the size of the input image using bilinear interpolation.

3.1.2. Evaluation Metrics. We use the three metrics to evaluate the proposed method: *F*-degree score, MAE, and *S*-degree score.

The calculation of *F*-degree F_β is as follows:

$$F_\beta = \frac{(1+\beta)\mathrm{TP}}{\left(1+\beta^2\right)+\beta^2\mathrm{FN}+\mathrm{FP}}, \tag{7}$$

where TP means that the classifier recognizes the correct positive sample; TN means that the classifier recognizes the correct negative sample; FP means the classifier recognizes the wrong negative sample; and FN means the classifier recognizes the wrong positive sample. Precision is defined as

$$P = \frac{TP}{TP+FP}. \tag{8}$$

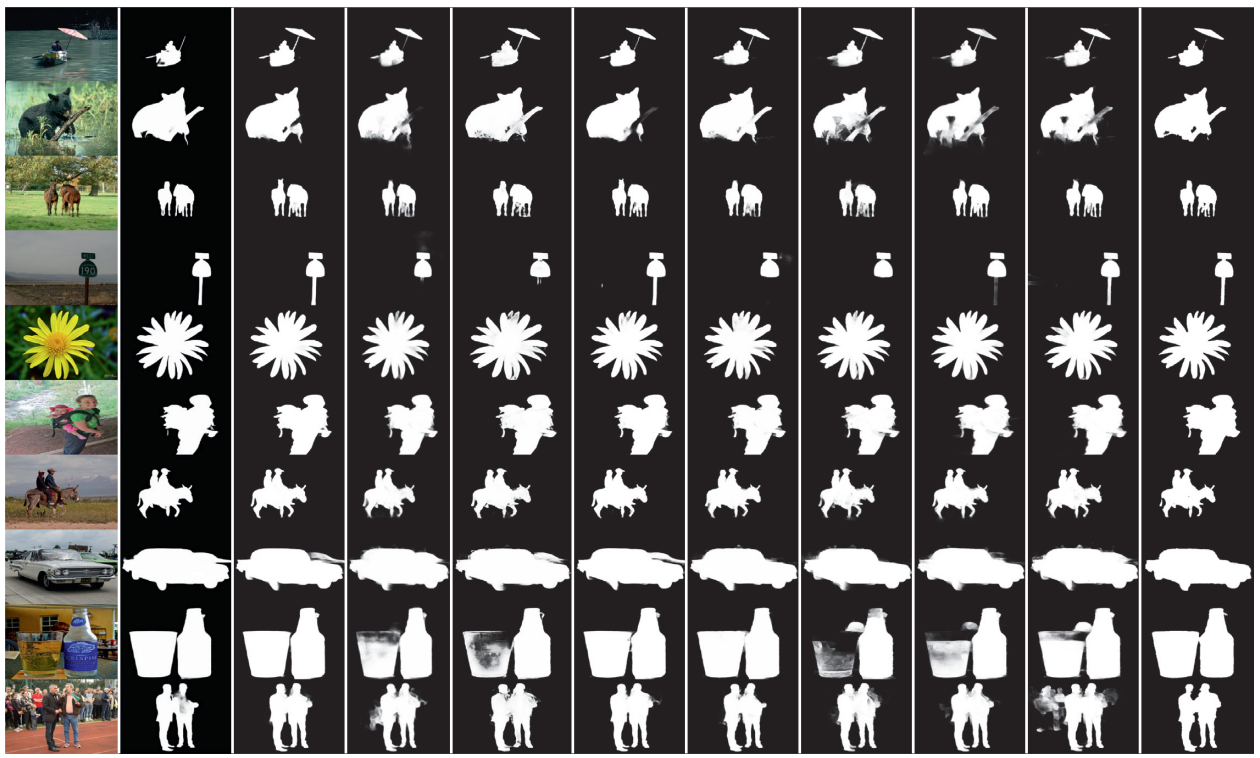
Recall is defined as

$$R = \frac{TP}{TP+FN}, \tag{9}$$

MAE represents the average of the absolute error between predicted and observed values, namely, the prediction between the significant mapping and its real mask average absolute deviation per pixel. The MAE is a linear score which means that the weight of all the individual differences in average is equal. As a supplement of the PR curve, it is calculated by the average absolute difference between the pixel significant value and the ground truth:

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|S_i - G_i\right|. \tag{10}$$

where m represents the area of significance mapping; S_i represents the probability map of significance of the pixel; and G_i represents the real value of the pixel. *S* measure takes



(a) image (b) BASNet (c) F3Net (d) GateNet (e) ITSD (f) LDF (g) MINet (h) AFNet (i) EGNet (j) PoolNet (k) Ours

Figure 10: Qualitative comparison of ten methods on the ECSSD dataset.

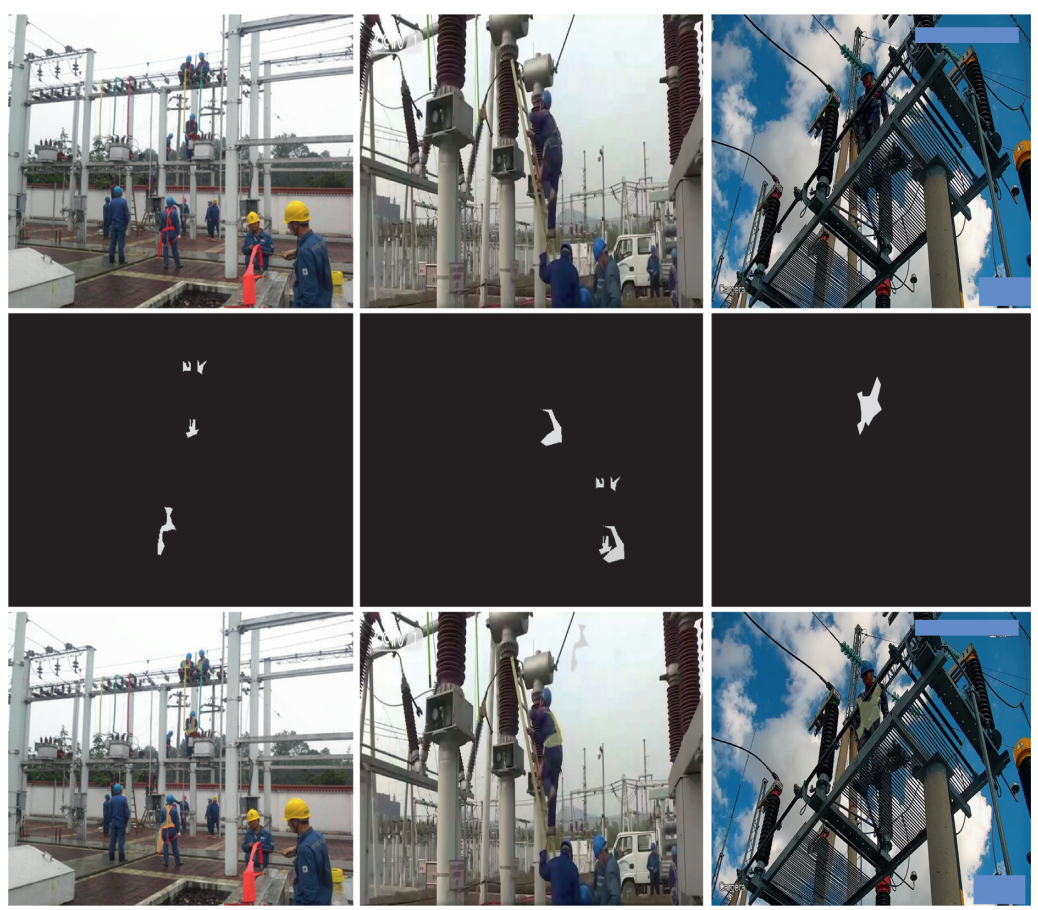


Figure 11: Experiments on safety harness. The first line shows the original images to be detected, the second line gives the detected saliency maps, and the third line demonstrates the matching effects between the saliency maps and the targets.

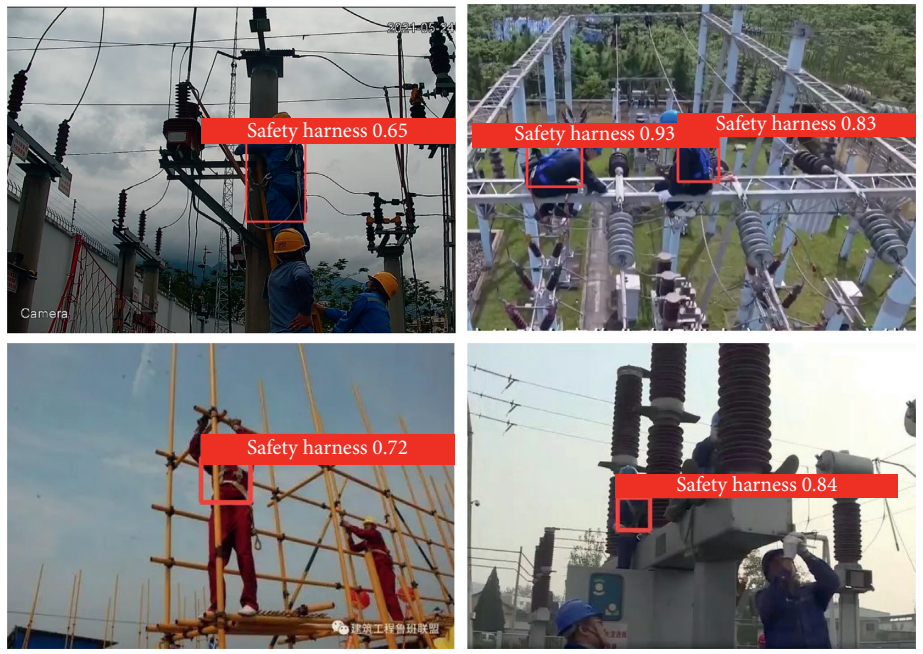


Figure 12: Detection results under the complex background using the proposed method.

into account the structural similarity of the area perception junction (SR) and object perception (SO), where α is set to 0.5.

$$S = \alpha * SO + (1 - \alpha) * SR \tag{11}$$

3.2. Ablation Study. In this section, we test the validity of each component proposed in our model and conduct ablation experiments on the ECSSD dataset. To demonstrate the effectiveness of our detection enhancement network, we first use the FPN branch, and the proposed IPU, DFBU, SRU, HMU, and FRU are then added in turn. Table 1 gives the results of this ablation study.

3.3. Quantitative Evaluation. We compare our model with nine other models: AFNet, BASNet, EGNet, F3Net, GateNet, ITSD, LDF, MINNet, and PoolNet. To evaluate the qualities of the segmented protruding objects, Table 2 summarizes the F measure (F_β), S measure (S_m), and MAE measure for the largest region of all datasets. As Table 2 shows, the proposed method outperforms other methods in both area and boundary measures by using ResNet-50 as a backbone. In particular, our method improves by 4.1%, 5.1%, 6.2%, 3.4%, and 5.9% on ECSSD, HKU-IS, DUT-OMRON, DUTS-TE, and PASCAL-S datasets, respectively.

To further demonstrate the superior performance of our method, we show a qualitative comparison of our method with other methods in Figure 10. We can see that the proposed method can suppress the interference information in the case of complex background and strengthen the effective information of saliency targets in images.

3.4. Practical Application. In order to solve the problem of safety harness detection in power production safety monitoring, we apply the proposed double U-shaped multi-reinforced unit structure network in the YOLOv5 detection model and test the performance on the aforementioned power construction site dataset. Figure 11 shows that the proposed method can accurately detect target saliency maps under the complex power construction site background and improve the detection accuracy by 10% compared with the original YOLOv5 network as shown in Figure 12.

4. Conclusions

In this paper, we have proposed a double U-shaped multi-reinforced unit structure network (DUMRN) to improve object detection. The proposed network consists of the detection module (DM), the reinforced module (RM), and the salient loss function (SLF). Quantitative evaluation on five public datasets has been conducted, and experimental results show that the proposed method gives the accurate performance and outperforms nine state-of-the-art methods. In addition, the safety harness detection experiment further verifies the effectiveness of the proposed method in the practical application. However, there are still some shortcomings in the proposed method. First of all, compared with general object detection methods, the proposed method consumes more time due to the salient object detection preprocess. Secondly, the proposed method may not provide a stable performance for small target detection. In the future, we will further expand datasets for more practical applications and improve the speed of the proposed method by optimizing network structures.

Data Availability

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

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors' Contributions

Qian Zhao and Haifeng Wang contributed equally to this work.

Acknowledgments

The authors thank for the dataset for safety harness detection provided by Yunnan Power Grid Co., Ltd., Yuxi Power Supply Bureau, and the openCODE provided by Xuebin Qin et al. This work was supported by the Yunnan Province Ten Thousand Talents Program, Postgraduate Research Innovation Fund Project of Yunnan Normal University (ysdyjs2020148), and Science and Technology Innovation Program of the Institute of Optics and Electronics, Chinese Academy of Sciences (20204001026).

References

[1] K. Fu, C. Gong, J. Yang, Y. Zhou, and I. Yu-Hua Gu, "Superpixel based color contrast and color distribution driven salient object detection," *Signal Processing: Image Communication*, vol. 28, no. 10, pp. 1448–1463, 2013.

[2] J. Han, D. Zhang, X. Hu, G. Lei, R. Jinchang, and W. Feng, "Background prior-based salient object detection via deep reconstruction residual," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 8, pp. 1309–1321, 2014.

[3] N. Singh, R. Arya, and R. K. Agrawal, "A novel position prior using fusion of rule of thirds and image center for salient object detection," *Multimedia Tools and Applications*, vol. 76, no. 8, pp. 10521–10538, 2017.

[4] R. S. Srivatsa and R. V. Babu, "Salient object detection via objectness measure," in *Proceedings of the 2015 IEEE international conference on image processing (ICIP). IEEE'2015*, pp. 4481–4485, Quebec City, QC, Canada, September 2015.

[5] C. Li, R. Cong, S. Kwong et al., "ASIF-Net: attention steered interweave fusion network for RGB-D salient object detection," *IEEE transactions on cybernetics*, vol. 51, no. 1, pp. 88–100, 2020.

[6] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: a benchmark," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5706–5722, 2015.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Vision and Pattern Recognition*, vol. 1409, Article ID 1556, 2014.

[9] C. F. Benitez-Quiroz, S. Rivera, P. F. U. Gotardo, and A. M. Martinez, "Salient and non-salient fiducial detection using a probabilistic graphical model," *Pattern Recognition*, vol. 47, no. 1, pp. 208–215, 2014.

[10] H. Zhang, T. Zhang, W. Pedrycz, C. Zhao, and D. Miao, "Improved adaptive image retrieval with the use of shadowed sets," *Pattern Recognition*, vol. 90, pp. 390–403, 2019.

[11] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *Proceedings of the European conference on computer vision*, pp. 825–841, Amsterdam, The Netherlands, October 2016.

[12] Z. Li, C. Lang, L. Liang et al., "Dense attentive feature enhancement for salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[13] P. Zhang, D. Wang, H. Lu, W. Hongyu, and Y. Baocai, "Learning uncertain convolutional features for accurate saliency detection," in *Proceedings of the IEEE International Conference on computer vision*, pp. 212–221, Venice, Italy, October 2017.

[14] P. Hu, B. Shuai, J. Liu, and W. Gang, "Deep level sets for salient object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2300–2309, Honolulu, HI, USA, July 2017.

[15] S. Osher and J. A. Sethian, "Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations," *Journal of Computational Physics*, vol. 79, no. 1, pp. 12–49, 1988.

[16] Z. Luo, A. Mishra, A. Achkar, E. Justin, L. Shaozi, and J. Pierre-Marc, "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.

[17] D. B. Mumford and J. Shah, "Optimal approximations by piecewise smooth functions and associated variational problems," *Communications on Pure and Applied Mathematics*, vol. 42, no. 5, 1989.

[18] Q. Hou, M. M. Cheng, X. Hu, B. Ali, T. Zhuowen, and T. Philip, "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.

[19] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 1395–1403, Santiago, Chile, December 2015.

[20] S. Chen, X. Tan, and B. Wang, "Reverse attention for salient object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, vol. 11213, pp. 234–250, Munich, Germany, September 2018.

[21] P. Zhang, W. Liu, H. Lu, and S. Chunhua, "Salient object detection by lossless feature reflection," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, vol. 1802, Article ID 06527, Stockholm, Sweden, July 2018.

[22] Z. Yao and L. Wang, "ERBANet: Enhancing region and boundary awareness for salient object detection," *Neurocomputing*, vol. 448, pp. 152–167, 2021.

[23] Z. Deng, X. Hu, L. Zhu et al., "R3net: recurrent residual refinement network for saliency detection," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 684–690, Stockholm, Sweden, July 2018.

[24] M. A. Islam, M. Rochan, S. Naha, N. D. B. Bruce, and W. Yang, "Gated Feedback Refinement Network for Coarse-To-Fine Dense Semantic Image Labeling," in *Computer Vision and Pattern Recognition*, pp. 1806–11266, World Scientific Publishing Company, Salt Lake City, UT, USA, 2018.

[25] X. Xu, J. Chen, H. Zhang, and G. Han, "Dual pyramid network for salient object detection," *Neurocomputing*, vol. 375, pp. 113–123, 2020.

[26] S. Zhou, J. Wang, J. Zhang et al., "Hierarchical U-shape attention network for salient object detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 8417–8428, 2020.

[27] A. Li, J. Qi, and H. Lu, "Multi-attention guided feature fusion network for salient object detection," *Neurocomputing*, vol. 411, pp. 416–427, 2020.

[28] X. Zhang, T. Wang, J. Qi, L. Huchuan, and W. Gang, "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, June 2018.

[29] X. Hu, L. Zhu, J. Qin, F. Chi-Wing, and H. Pheng-Ann, "Recurrently aggregating deep features for salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, New Orleans, LO, USA, January 2018.

[30] L. Wang, L. Wang, H. Lu, Z. Pingping, and R. Xiang, "Salient object detection with recurrent fully convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1734–1746, 2018.

[31] 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.

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

[33] T. Wang, A. Borji, L. Zhang, Z. Pingping, and L. Huchuan, "A stagewise refinement model for detecting salient objects in images," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4019–4028, Venice, Italy, October 2017.

[34] I. M. d. Amirul, Y. Wang, M. Kalash, M. Rochan, and N. D. B. Bruce, "Salient object detection using a context-aware refinement network," in *Proceedings of the British Machine Vision Conference 2017*, London, England, September 2017.

[35] K. Fu, Q. Zhao, and I. Y. H. Gu, "Refinet: A deep segmentation assisted refinement network for salient object detection," *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 457–469, 2018.

[36] X. Qin, Z. Zhang, C. Huang, G. Chao, D. Masood, and J. Martin, "BASNet: Boundary-Aware Salient Object Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7479–7489, Seoul, Korea, November 2019.