Improved High-Resolution Salient Object Detection Algorithm Based on Enhanced PGNet

Z. H. Wang, Y. Xu

Abstract—Significant progress has been made in salient object detection (SOD) based on deep neural networks. However, existing SOD methods are primarily designed for low-resolution image inputs, which often suffer from issues such as sampling depth, receptive field, and model performance when applied to high-resolution image inputs. To tackle this issue, this study introduces a lightweight feature extraction model by replacing the original ResNet network structure with a RepVGG-based model that incorporates the Efficient Channel Attention (ECA) module for lightweight purposes. In order to enhance both the model's accuracy and processing speed, we introduce the Effective Squeeze-and-Excitation (ESE) module for feature fusion. To tackle the issue of unclear boundaries of salient objects, we fuse the Weighted Binary Cross-Entropy, Structural Similarity (SSIM), and Shape-aware Loss into a combined loss function, which replaces the conventional cross-entropy loss. Experimental results demonstrate that the enhanced algorithm (RepPGNet) achieves a 2.8% increase in accuracy compared to the original algorithm, with reduced model parameters and improved clarity of salient object boundaries. The proposed algorithm is also shown to have improved speed and is suitable for high-resolution image scenes.

***Index Terms*—Saliency detection, High-resolution, PGNet, Computer vision**

I. Introduction

Saliency detection[1] refers to the process of identifying the most attention-grabbing objects within prominent scenes. Over the recent years, research on saliency detection has grown significantly, leading to its wide-ranging applications in image classification, image compression, image segmentation[2], object detection, autonomous driving, visual tracking, among others.

Over the past few decades, a multitude of traditional methods have emerged to tackle the challenges of saliency detection. Nevertheless, these approaches predominantly emphasize low-level features, often neglecting the wealth of semantic information available, leading to inconsistent performance, particularly in intricate scenes. In recent years, substantial advancements in saliency object detection have been achieved through the application of deep neural networks. Hou et al. [3] utilized deep convolutional networks as encoders to extract multi-level features and designed multiple modules for feature fusion in a Feature Pyramid Network (FPN) style. Additionally, Wei et al. [4] achieved saliency maps with clear boundaries through explicit supervision the generation of edge pixels. The widespread use of Transformers [5] in visual tasks has also brought new advancements in saliency detection. Nevertheless, the majority of these approaches are tailored for low-resolution environments and do not readily extend to high-resolution scenarios.

The majority of existing saliency detection methods perform well within fixed low-resolution input ranges, such as 224 × 224 or 384 × 384. Due to the rapid advancement of image capture devices, the resolutions of the images available to us, such as 1080p, 2K, and 4K, far exceed the range that existing saliency detection methods can directly adapt to. As a result, many algorithms have emerged to address high-resolution image challenges. Guo[6] and colleagues proposed a new super-resolution segmentation framework called ISDNet, which integrates shallow and deep networks in a novel way, effectively addressing the significant computational and memory burden. Shen et al.[7] introduced the Continuous Refinement Model (CRM) to bridge the resolution disparity between low-resolution training images and high-resolution test images. Additionally, Xie[8] and collaborators presented a single-stage pyramid grafting network (PGNet) that fuses fragmented information through cross-model grafting modules. Research findings have demonstrated that the PGNet model exhibits outstanding performance in terms of speed, accuracy, and various other aspects when compared to state-of-the-art algorithms concurrently. Although PGNet outperforms other algorithms in various aspects, there is still potential for enhancing the capability to detect prominent object boundaries. Therefore, it is meaningful to make improvements on PGNet.

Given the issues such as the large size and slow speed of PGNet, this study proposes an improved high-resolution significant object detection algorithm based on PGNet, which replaces the original backbone network ResNet with a more lightweight model for feature extraction, reduces the parameter count while improving the model's speed and feature focus. Finally, by adjusting the loss function, the sharpness of significant object boundaries is further enhanced.

Manuscript received April 25, 2023;revised September 8, 2023.This work was supported by the National Natural Science Foundation of China (61775169), the Education Department of Liaoning Province (LJKZ0310), the Excellent Young Talents Program of Liaoning University of Science and Technology (2021YQ04).

Z. H. Wang is a postgraduate student at the School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan 114051, China (e-mail: 1057476776@qq.com).

Y. Xu is a Professor at the School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan 114051, China (corresponding author, phone: 86-13889785726; e-mail: xuyang_1981@aliyun.com).

II. Principle of the PGNet Algorithm

The architecture of PGNet is illustrated in Fig.1, where two encoders, namely Swin Transformer and ResNet-18, are parallelly employed. The Transformer encoder captures accurate global semantic information from low-resolution image inputs, while the ResNet encoder extracts rich detailed information from high-resolution image inputs. The information from both encoders can complement each other. The features obtained from the Transformer encoder are then grafted onto the ResNet branch through the proposed Cross-Model Grafting Module (CMGM), as shown in Fig.2, to achieve higher-level pyramid features (by grafting two lower-resolution feature maps from the Transformer encoder), with minimal computational cost and friendliness to high-resolution image inputs. Unlike common feature fusion strategies that rely on pixel-wise operations such as addition or multiplication, which are limited to local information and prone to errors, the objective of CMGM is to graft Transformer information into the ResNet branch. This is achieved by recalculating ResNet and Transformer features point-wise, transferring global semantic information from the Transformer encoder to the ResNet branch. Additionally, PGNet incorporates an Attention-Guided Loss to supervise the Cross Attention Matrix (CAM).

In summary, PGNet utilizes an interleaved connection structure to build a higher-level feature pyramid from two lower-level pyramids and proposes the Cross-Model Grafting Module to graft features extracted from two encoders. In essence, the network achieves greater sampling depth while minimizing computational costs, addressing the demands presented by high-resolution inputs.

III. Improved Strategy

The manuscript presents optimizations in two aspects based on various metrics of high-resolution saliency detection and the PGNet algorithm itself, namely, optimization of the network backbone and adjustment of the loss function. The backbone of PGNet adopts ResNet and Transformer, which is hierarchical and densely connected, capable of extracting deep features effectively. However, it may suffer from redundancy, resulting in a substantial increase in the number of parameters and an inflated model size. Therefore, in this study, a lightweight model that is faster, memory-efficient, and more flexible is chosen to replace ResNet to reduce the model's parameter count and improve its speed and accuracy. The first part of the optimization focuses on the network backbone, which serves as the fundamental feature extractor for the detection task, responsible for extracting salient features from input images and producing output. The lightweight feature extraction model proposed in this study is mainly based on RepVGG[9], with the incorporation of the ECA[10] module to achieve lightweight effects, and the addition of the ESE[11] module for feature fusion, thereby improving the accuracy and speed of the model. The second part entails modifying the loss function, where the original conventional standard cross-entropy loss is improved by integrating a mixed loss function that combines weighted binary cross-entropy loss, structural similarity index (SSIM), and shape-aware loss.

This adjustment effectively enhances the boundary clarity of the saliency map. The improved model, named RepPGNet, is illustrated in Fig.3, where (a)(b)(c) represent the decoder blocks in the decoding module, and DBn (Decoder block) corresponds to the decoding module with n input features (n=1, 2, 3).

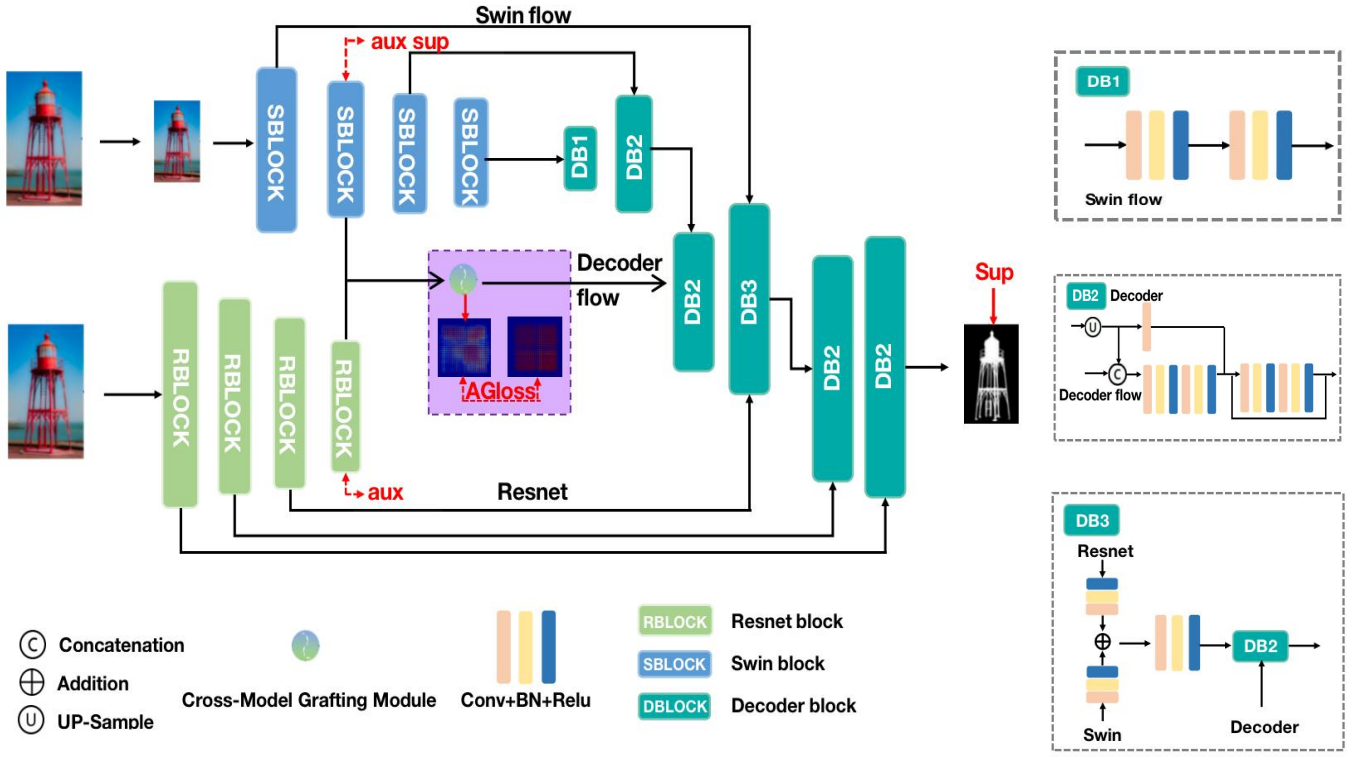


Fig. 1. Illustrates the basic structure of PGNet.

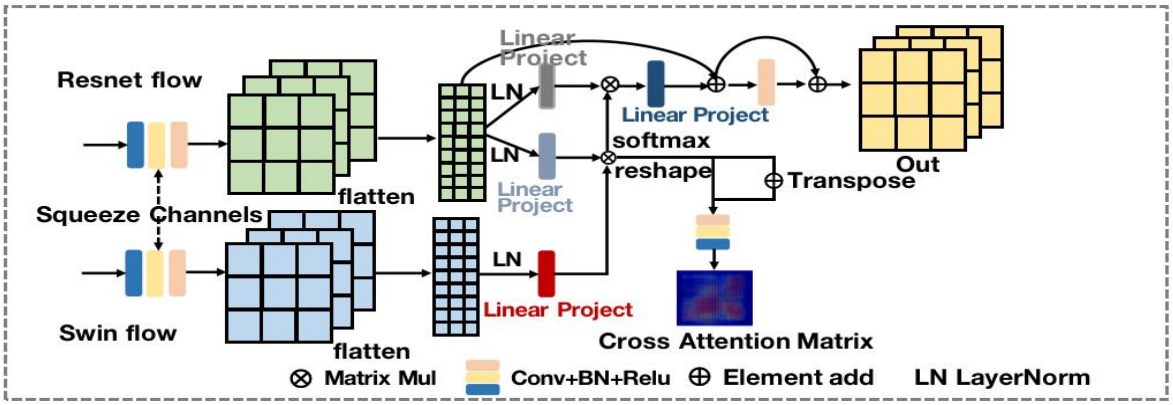


Fig. 2. The architecture of cross model grafting module.

A. *Optimization of Backbone Networks*

RepVGG is notably the initial standard model to attain a Top-1 accuracy exceeding 80% on the ImageNet dataset, particularly when utilizing an NVIDIA 1080Ti GPU, the RepVGG model exhibits an 83% speed improvement over ResNet-50 [12] and a 101% speed enhancement over ResNet-101, while maintaining superior accuracy in comparison to state-of-the-art models such as EfficientNet [13] and RegNet [14]. RepVGG is a simple yet powerful convolutional neural network architecture, with a similar inference time as VGG [15], consisting of only multiple 3×3 convolutions and ReLU activation functions, while the model during training exhibits a multi-branch topology. The accuracy and speed are achieved by decoupling the training process and inference time, using parameter fusion during forward inference, a technique called reparameterization.

Fig.4 (A) depicts the original ResNet network, which contains residual structures of Conv1×1 (1×1 convolution) and Identity activation function, solving the gradient vanishing problem in deep networks and making the network more prone to convergence. Figure (B) illustrates the training phase architecture of the RepVGG network, with an overall structure akin to ResNet, featuring residual structures in both. The main difference between the two networks is that the residual blocks in the RepVGG network do not skip across layers, and as the model goes deeper, more complex residual structures are used, which not only allows for more robust feature representation in deep layers but also better handles the gradient vanishing problem in deep layers. Figure (C) represents the inference phase architecture of RepVGG network, which is very simple, consisting of only 3 × 3 convolutions and ReLU activations, making it efficient for model inference and acceleration.

The benefits of RepVGG stem from its utilization of distinct network architectures for both training and inference stages. During the training phase, the emphasis is on achieving accuracy, whereas in the inference phase, the primary concern is speed.

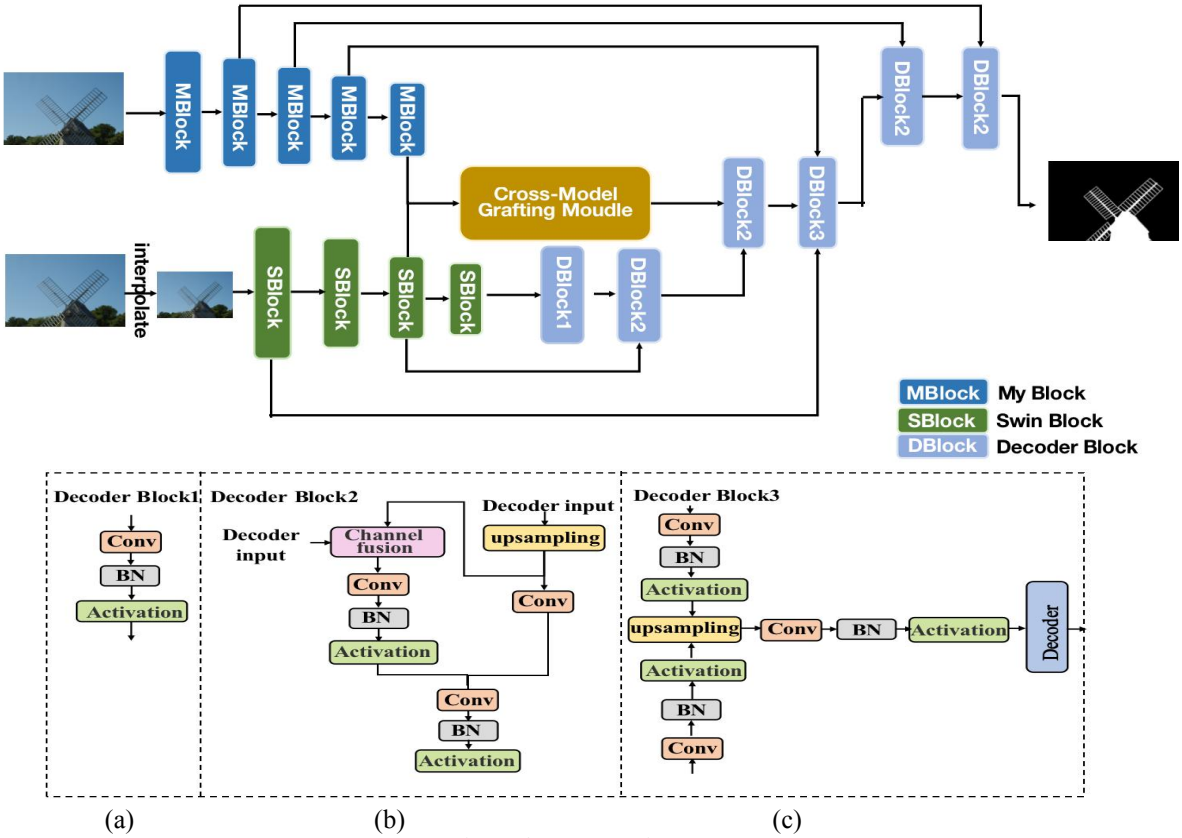


Fig. 3. The structure of RepPGNet.

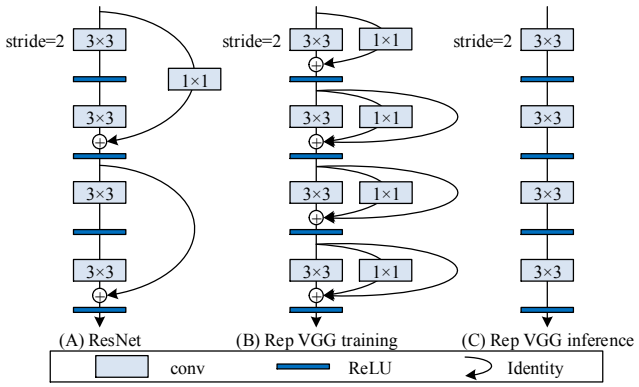


Fig. 4. Schematic diagram of RepVGG architecture.

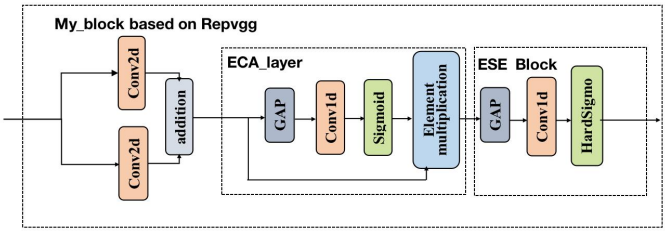


Fig. 5. Lightweight feature extraction model.

The main purpose of this lightweight model is to perform feature extraction, which is based on RepVGG and incorporates ECA and ESE for achieving a more lightweight design, as shown in Fig.5.

1. Efficient Channel Attention Module (ECA): ECA is an extremely lightweight and plug-and-play attention module that improves the performance of various deep convolutional neural network (CNN) architectures. It contains only a small number of parameters but brings significant performance improvement. The ECA module adopts a non-reducing local cross-channel interaction strategy, effectively avoiding the impact of dimension reduction on channel attention learning. The idea of the ECA module is very simple, as shown in Fig.6:

(1) Local inter-channel interaction in one-dimensional convolution is achieved through fast convolution with a kernel size of k, where the kernel size k represents the coverage range of inter-channel interaction, i.e., the number of regions involved in the attention prediction of a channel.

(2) To avoid manual adjustment of k through cross-validation, a self-adaptive method has been developed to determine k, where the coverage range of inter-channel interaction (i.e., kernel size k) is proportional to the channel dimension.

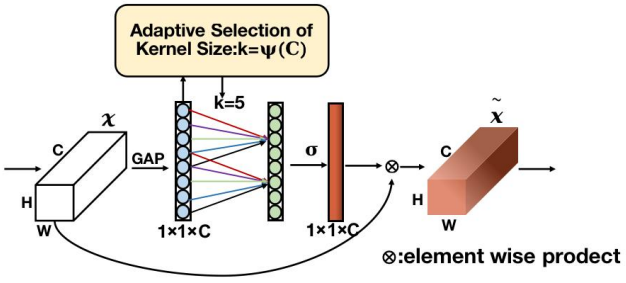


Fig. 6. The structure of ECA.

2. ESE: In this study, an Effective Squeeze-Excitation (ESE) module, which is a more efficient improvement of the original SE, is employed. SE is a representative channel attention method used in CNN architectures to explicitly model the interdependencies among feature map channels for enhanced representation capability. The SE module compresses the spatial correlations through global average pooling to learn descriptors for specific channels, and then rescales the input feature map using two fully connected (FC) layers and a sigmoid function to highlight informative channels. In short, given an input feature map $X_i \in \mathbb{R}^{C\times W\times H}$, the computation formula for the channel attention map $A_{ch}(X_i) \in \mathbb{R}^{C\times 1\times 1}$ refer to"(1)".

$$A_{ch}(X_i) = \sigma(W_C(\delta(W_{C/r}(F_{gap}(X_i))))) \quad (1)$$

In this context, C represents channel-wise global average pooling, D denotes the weights of two fully connected (FC) layers, δ represents the ReLU activation function, and σ represents the sigmoid function. However, the Squeeze-and-Excitation (SE) module has limitations, as the reduction in dimensionality leads to channel information loss. To alleviate the complexity associated with model expansion, the SE module incorporates two Fully Connected (FC) layers. The primary FC layer reduces the input feature channels from C to C/r, employing a compression ratio denoted as 'r.' Subsequently, the secondary FC layer restores the compressed channels to their original size of C. As a result, the reduction in channel dimensionality leads to channel information loss. Therefore, this study proposes an efficient version of the SE module, referred to as Efficient Squeeze-and-Excitation (ESE), which employs only one FC layer with C channels instead of two FC layers, avoiding channel information loss and improving performance.

B. Optimization of the Loss Function

Currently, most loss functions used in salient object detection rely on cross-entropy (CE) loss. However, due to its limitations in differentiating boundary pixels during training, it may result in blurred boundaries. To obtain saliency maps with high confidence and clear boundaries, this study proposes a hybrid loss function by modifying the traditional CE loss, which combines weighted binary cross-entropy (Weighted BCE) [16], structural similarity index (SSIM) [17], and shape-aware loss [18].

1.Weighted Binary Cross-Entropy refer to"(2)".

$$WCE(p,\hat{p}) = -(\beta p\log(\hat{p}) + (1-p)\log(1-\hat{p})) \quad (2)$$

The weighted binary cross-entropy (BCE) loss function is an extension of the original cross-entropy loss by incorporating weight parameters for each class. By setting a weight parameter $\beta > 1$, false negatives can be reduced, while setting $\beta < 1$ can reduce false positives. This approach improves the performance of the loss function in scenarios where sample imbalance exists, compared to the original cross-entropy loss.

2. The structural similarity (SSIM) index was originally proposed for image quality assessment. Owing to the robust correlations among pixels within an image, these correlations encapsulate crucial information regarding object structures within visual scenes. SSIM defines the structural information of an image based on the brightness and contrast related to the object structure, allowing it to capture the structural information in the input image.

SSIM measurement system comprises three contrast modules, namely brightness, contrast, and structure.

The estimation of brightness measurement requires the calculation of the image's average grayscale level using a brightness contrast function. Refer to"(3)".

$$\mu_X = \frac{1}{H \times M} \sum_{i=1}^{H} \sum_{j=1}^{M} X(i,j) \tag{3}$$

Therefore, which represents the brightness contrast function for two images. Refer to"(4)".

$$l(x,y) = \frac{2\mu_x \mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{4}$$

The contrast function determines the image's standard deviation by applying the contrast measurement estimation formula refer to "(5)".

$$\sigma_X = \left(\frac{1}{H+W-1} \sum_{i=1}^{H} \sum_{j=1}^{M} (X(i,j) - \mu_X)^2\right)^{\frac{1}{2}} \tag{5}$$

Thus, the contrast function of the two images refer to"(6)".

$$c(x,y) = \frac{2\sigma_x \sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{6}$$

The structural comparison function refer to"(7)".

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x \sigma_y + C_3} \tag{7}$$

Finally, the formula for the SSIM function is obtained by integrating the three comparison functions mentioned above. refer to"(8)".:

$$\begin{aligned} SSIM(x,y) &= f\left(l(x,y), c(x,y), s(x,y)\right) \\ &= \left[l(x,y)\right]^{\alpha} \left[c(x,y)\right]^{\beta} \left[s(x,y)\right]^{\gamma} \end{aligned} \tag{8}$$

where α, β, and γ, all of which are greater than 0, serve to fine-tune the significance of the three modules. If α, β, and γ are all 1, then $C_3 = \frac{C_2}{2}$,refer to"(9)".

$$SSIM(x,y) = \frac{(2\mu_x \mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{9}$$

3.The Shape-aware Loss considers the shape information in addition to pixel-level losses commonly used in other loss functions. Specifically, it calculates the average Euclidean distance between the predicted segmentation curve and the ground truth curve at points surrounding the curve, and uses this distance as a coefficient for the cross-entropy (CE) loss function. The formulation of the Shape-aware Loss refer to " (10) " .(Where "CE" refers to the cross-entropy loss functionrefer to"(11)")

$$L_{shape-aware} = -\sum_{i} CE(y, \hat{y}) - \sum iE_i CE(y, \hat{y}) \tag{10}$$

$$E_i = D(\hat{c}, C_{GT}) \tag{11}$$

The research findings indicate that employing a hybrid loss function as described may yield improved details and edges, effectively enhancing the overall accuracy of salient region detection.

IV. Analysis of Experiments And Results

A. Dataset Selection

(1) High-resolution datasets: The available high-resolution datasets used in this study include UHRSD [19] (4932 images for training and 988 images for testing), which contains diverse and rich scenes with significant objects of various sizes, complexities, and levels of detail. The HRSOD dataset [20] (1610 images for training and 400 images for testing) is also used in this study. Additionally, DAVIS-S [21], which is a dataset with pixel-perfect annotated object masks, is used for evaluation.

(2) Low-resolution dataset: DUTS-TR [22], which is part of the DUTS dataset and contains 10553 images, is the largest and most commonly used training dataset for salient object detection, and is used for training the model in this study. Furthermore, the proposed method is evaluated on widely used benchmark datasets, including ECSSD [23] with 1000 images, DUT-OMRON [24] with 5168 images, PASCAL-S [25] with 850 images, DUTS-TE [22] with 5019 images, and HKU-IS [26] with 4447 images.

B. Experimental Platform and Resource Consumption

The experiments were carried out on a computer system featuring an Intel Core i7-11700 CPU and an NVIDIA GeForce GTX 3070 GPU. The experiments were conducted on a system running the Windows 10 operating system, utilizing the PyTorch 1.8-GPU deep learning framework, and the PyCharm Community IDE. During the training process, the model employed the mini-batch gradient descent algorithm with a batch size of 32, running 1200 iterations on the training set. Each iteration took approximately 1.1 minutes and consumed 8GB of GPU memory.

During the inference process, the model performed saliency detection on a high-resolution image with an average processing time of 0.09 seconds and used 0.5GB of GPU memory. Since offline training was used and the network bandwidth requirements during the testing phase were negligible, network bandwidth consumption was not a consideration.

C. Experimental Evaluation Criteria

In this study, the performance of the algorithm is evaluated using the following metrics. Firstly, the Mean Absolute Error (MAE), refers to(12), where S represents the predicted image, G represents the ground truth, and W and H denote the width and height of S, respectively.

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \left| S(x,y) - G(x,y) \right| \tag{12}$$

The second metric is the Max F-measure (F_{β}^{Max}), where β2 is set to 0.3, refer to (13),

$$F_{\beta} = \frac{(1+\beta^2) \times \mathrm{Precision} \times \mathrm{Recall}}{\beta^2 \times \mathrm{Precision} + \mathrm{Recall}} \tag{13}$$

Finally, in this study, the Structural Similarity (S-measure) is employed as a performance evaluation metric, refer to(14). The E-measure (Eξ) can be referred to in [27].

$$S = \alpha \times S_o + (1-\alpha) \times S_r \tag{14}$$

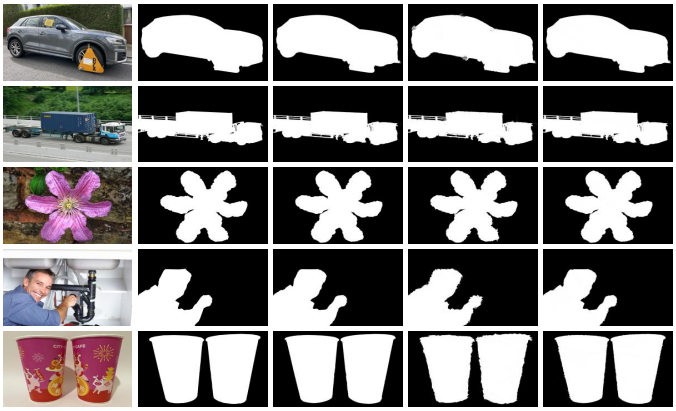
In this study, the value of α is configured at 0.5, where So signifies the similarity of object structures, and Sr denotes the similarity of region structures.

D. Ablation Experiments

The effectiveness of the loss function is demonstrated through the introduction of a lightweight feature extraction model and the optimization of the network structure. Table I presents the results of ablation experiments for data comparison. Subsequently, a comparison of the optimized parameter quantity and network model size is shown, followed by a comparative analysis of the experimental results against other models used for salient object detection. To validate the effectiveness of RepPGNet, this study conducts experiments on each optimization scheme separately during the optimization process. Table I reveals that RepPGNet attains the highest level of performance.

TABLE I
COMPARISON OF ABLATION EXPERIMENTAL DATA

Compsition	HRSOD-TE			
	F_β^{Max}	MAE	$E\xi$	Sm
Baseline-ResNet+RepVGG	0.931	0.026	0.937	0.928
Baseline-ResNet+RepVGG +ECA	0.942	0.023	0.948	0.939
Baseline-ResNet+RepVGG +ECA+ESE+LOSS(ours)	**0.966**	**0.020**	**0.977**	**0.950**



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

Fig. 7. Comparison of ablation experiment effects.

As illustrated in Figure 7, (a) represents the input image, (b) is the ground truth of the image, (c) shows the results obtained from RepPGNet, (d) shows the results obtained by replacing only the backbone network with RepVGG (Rep), and (e) shows the results obtained by incorporating the ECA attention mechanism into RepVGG (Rep+ECA). It is evident from the comparison in (d) that the saliency map obtained after replacing ResNet is not as satisfactory. This is attributed to the significant reduction in parameters in RepVGG, which eliminates redundancy in the model data but sacrifices some features. However, from the comparative analysis of the results in (c), (d), and (e), it can be observed that RepPGNet effectively addresses this issue.

E. Implementation details and Parameters

In this study, we employed the mini-batch Adam algorithm for model optimization and initialized the parameters using the Kaiming method. This optimization algorithm exhibits efficient computational performance and excellent handling of sparse gradients. To address scale variations in images, we utilized bilinear convolution as the upsampling and downsampling function. To prevent overfitting, we introduced Batch Normalization and Dropout techniques. Batch Normalization ensures that the inputs to each layer follow a standard normal distribution, while Dropout temporarily drops out some neurons during the training process. Grid search was employed to select the optimal hyperparameter values. For the Dropout parameter, we explored values in the range of [0.1, 0.2, 0.3] and ultimately determined the optimal value to be 0.1. Furthermore, we incorporated a Learning Rate Scheduler to implement an adaptive learning rate decay strategy, further optimizing the model's performance.

Observing Table II, it becomes apparent that RepPGNet exhibits the lowest parameter count when compared to the original PGNet model.

TABLE II
COMPARISON OF NETWORK MODEL SIZES

Model	Parameters
PGNet-ResNet	72,666,404
RepPGNet	**63,347,827**

To highlight the superiority of RepPGNet in terms of inference speed, this study conducted a comparison between RepPGNet and PGNet. To ensure consistency and eliminate randomness, this comparison was carried out on a standardized set of 500 images, each with resolutions of 1024 ×1024, 2k, and 4k. The average inference speed, measured in images processed per second, was calculated for each image. The results, as presented in Table III, clearly demonstrate that RepPGNet outperforms PGNet in terms of inference speed.

TABLE III
COMPARISON OF REASONING SPEED BETWEEN REPPGNET AND PGNET

Model	1024×1024	2k	4k
RepPGNet	0.074795	0.852847	0.099844
PGNet	**0.083923**	**0.887194**	**0.102988**

To evaluate the efficacy of high-resolution salient object detection, this study conducts a quantitative performance comparison between RepPGNet and contemporary SOD models. This assessment encompasses five benchmark datasets, with evaluation metrics including max F-measure, MAE, E-measure, and S-measure. As illustrated in Table IV, the results demonstrate that RepPGNet exhibits a notable performance enhancement on the majority of high-resolution datasets when compared to competing algorithms.

To further exemplify the advantages of the RepPGNet approach, this study compares RepPGNet with nine state-of-the-art algorithms, including CTD, CPD, PGNet, DASNet, F3Net, GCPA, LDF, PFS, and SCRN. As shown in Table IV, RepPGNet has achieved significant improvements in high-resolution datasets.

To visually showcase the superiority of RepPGNet, Fig. 8 presents representative examples of visual comparisons between RepPGNet and other algorithms. Lines one to four are from the UHRSD-TE dataset, while lines five and six are from the HRSOD-TE dataset. It can be observed that RepPGNet is capable of capturing fine details and generating clear object boundaries (as shown in the first and second rows).

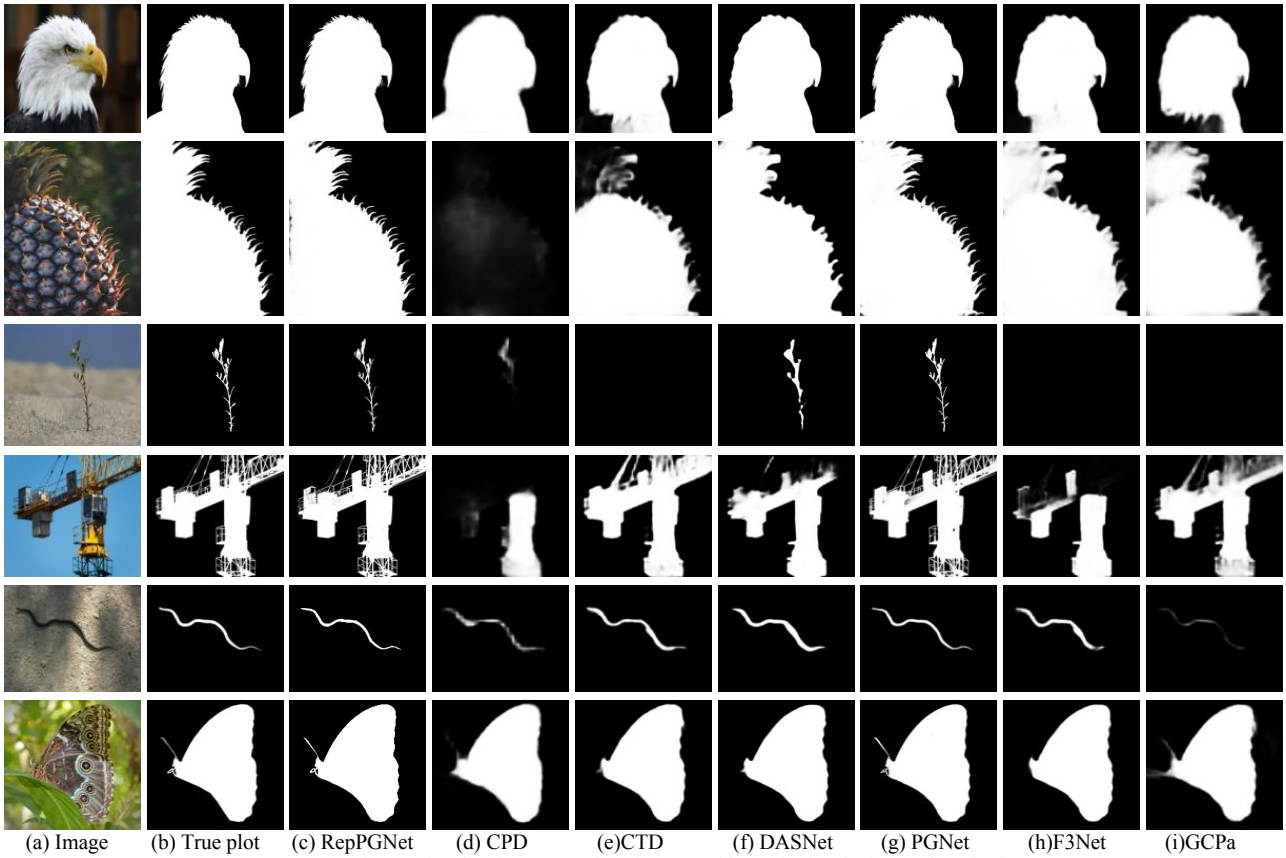
TABLE IV
PERFORMANCE COMPARISON OF VARIOUS METHODS

method	HRSOD-TE				DAVIS-S				UHRSD-TE				DUT-OMRON				DUTS-TE			
	F_β^{Max}	MAE	$E\xi$	Sm	F_β^{Max}	MAE	$E\xi$	Sm	F_β^{Max}	MAE	$E\xi$	Sm	F_β^{Max}	MAE	$E\xi$	Sm	F_β^{Max}	MAE	$E\xi$	Sm
SCRN[32]	0.880	0.042	0.887	0.888	0.893	0.027	0.911	0.902	0.904	0.051	0.880	0.887	0.811	0.056	0.863	0.837	0.888	0.040	0.888	0.885
CPD[29]	0.867	0.041	0.891	0.881	0.871	0.029	0.921	0.893	0.894	0.055	0.884	0.878	0.797	0.056	0.866	0.825	0.865	0.043	0.887	0.869
DASNet[30]	0.893	0.032	0.925	0.897	0.902	0.020	0.949	0.911	0.914	0.045	0.892	0.889	0.827	0.050	0.877	0.845	0.895	0.034	0.908	0.894
LDF[16]	0.904	0.032	0.919	0.904	0.911	0.019	0.947	0.922	0.913	0.047	0.891	0.888	0.820	0.051	0.873	0.838	0.898	0.034	0.910	0.892
F3Net[31]	0.900	0.035	0.913	0.897	0.915	0.020	0.940	0.914	0.909	0.046	0.887	0.890	0.813	0.053	0.871	0.838	0.891	0.035	0.902	0.888
GCPA[13]	0.889	0.036	0.898	0.898	0.922	0.020	0.934	0.929	0.912	0.047	0.886	0.896	0.812	0.056	0.860	0.839	0.888	0.038	0.891	0.891
PFS[33]	0.911	0.033	0.922	0.906	0.916	0.019	0.946	0.923	0.918	0.043	0.896	0.897	0.823	0.055	0.875	0.842	0.896	0.036	0.902	0.892
CTD[28]	0.905	0.032	0.921	0.905	0.904	0.019	0.938	0.911	0.917	0.043	0.898	0.897	0.826	0.052	0.875	0.844	0.897	0.034	0.909	0.893
PGNet[8]	0.931	0.021	0.944	0.930	0.936	0.015	0.947	0.935	0.931	0.037	0.904	0.912	0.835	0.045	0.887	0.855	0.917	0.027	0.922	0.911
RepPGNet	**0.959**	0.020	**0.977**	**0.950**	**0.960**	0.014	**0.978**	0.945	**0.958**	0.032	0.927	0.920	**0.846**	0.040	0.898	0.866	0.917	0.022	0.920	0.918

Besides accurate delineation of high-quality boundaries, a crucial facet of high-resolution SOD is its capability to segment tiny and nuanced objects that can be easily missed in low-resolution scenarios (as evident in the third, fifth, and sixth rows), further underscoring RepPGNet's superiority. Moreover, RepPGNet performs well even in extremely complex scenes (as shown in the fourth row). From the images, it's evident that RepPGNet excels at precisely identifying salient objects across a range of application scenarios, particularly in cases where the saliency contrast with the background is minimal, surpassing other algorithms in performance.

V. CONCLUSION

This study introduces an advanced saliency object detection method for high-resolution images, built upon an enhanced PGNet network. The method replaces the ResNet branch in PGNet with a lightweight feature extraction model and introduces an efficient loss function optimization scheme. Additionally, it leverages the cross-model grafting module and attention-guided loss proposed by PGNet, synergizing the advantages of ResNet and Transformer while compensating for their common shortcomings. The enhanced algorithm, named RepPGNet, achieves a 2.8% improvement in accuracy compared to the original PGNet, with reduced model size and increased speed. However, saliency detection on high-resolution images still faces various challenges, such as complex textures, intricate details, and a vast amount of information. To further enhance the accuracy of high-resolution image saliency detection models, various techniques can be employed, including segmentation and fusion, pyramid structures, context modeling, introducing prior knowledge, and multimodal information fusion. These methods enhance the model's saliency detection by considering image context and domain-specific knowledge, improving accuracy and robustness. Future research will delve deeper in this direction.



(a) Image (b) True plot (c) RepPGNet (d) CPD (e)CTD (f) DASNet (g) PGNet (h)F3Net (i)GCPa

Fig. 8. Visualization comparison between this method and advanced methods.

REFERENCES

[1] X. X. Yuan, and Y. Xu, "Salient Object Detection Based on Improved PoolNet," *Engineering Letters*, vol. 30, no.4, pp1558-1565, 2022.

[2] Qingrui Zhang, Mingqiang Yang, Kidiyo Kpalma, Qinghe Zheng, and Xinxin Zhang, "Segmentation of Hand Posture against Complex Backgrounds Based on Saliency and Skin Colour Detection," *IAENG International Journal of Computer Science*, vol. 45, no.3, pp435-444, 2018.

[3] Hou Q, Cheng M M, Hu X, et al., "Deeply supervised salient object detection with short connections," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3203-3212, 2017.

[4] Wei J, Wang S, Wu Z, et al., "Label decoupling framework for salient object detection," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13025-13034, 2020.

[5] Ze Liu, Lin T, Cao Y, et al., "Swin transformer: Hierarchical vision transformer using shifted windows," arXiv preprint arXiv:2103. 14030, 2021.

[6] S. Guo, et al., "ISDNet: Integrating Shallow and Deep Networks for Efficient Ultra-high Resolution Segmentation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4351-4360, 2022.

[7] Shen T, Zhang Y, Qi L, et al., "High Quality Segmentation for Ultra High-resolution Images," arXiv e-prints, 2021. DOI:10. 48550/arXiv. 2111. 14482.

[8] C. Xie, C. Xia, M. Ma, et al., "Pyramid Grafting Network for One-Stage High Resolution Saliency Detection," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11707-11716, 2022.

[9] X. Ding, X. Zhang, N. Ma, et al., "RepVGG: Making VGG-style ConvNets Great Again," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13728-13737, 2021.

[10] Wang Q, Wu B, Zhu P, et al., "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[11] Y. Lee, J. Park, "CenterMask: Real-Time Anchor-Free Instance Segmentation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13903-13912, 2020.

[12] He K, Zhang X, Ren S, et al., "Deep residual learning for image recognition," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016.

[13] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. arXivpreprint arXiv:1905.11946, 2019.

[14] Radosavovic I, Kosaraju R P, Girshick R, et al., "Designing network designspaces," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10428-10436, 2020.

[15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[16] Xie S, Tu Z, "Holistically-Nested Edge Detection," *International Journal of Computer Vision*, pp. 3-18, 2015.

[17] Wang Z, Simoncelli E P, Bovik A C, "Multiscale structural similarity for image quality assessment," *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, pp. 1398-1402, 2003.

[18] Liu J, Desrosiers C, Zhou Y, "Semi-supervised Medical Image Segmentation Using Cross-Model Pseudo-Supervision with Shape Awareness and Local Context Constraints," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Cham, 2022.

[19] Liu N, Zhang N, Wan K, et al., "Visual saliency transformer," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4722-4732, 2021.

[20] Wang Z, Simoncelli E P, Bovik A C, "Multiscale structural similarity for image quality assessment," *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, pp. 1398-1402, 2003.

[21] Zhong Y, Li B, Tang L, et al., "Disentangled high quality salient object detection," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3580-3590, 2021.

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

[23] Yan Q, Xu L, Shi J, et al., "Hierarchical saliency detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1155-1162, 2013.

[24] Yang C, Zhang L, Lu H, et al., "Saliency detection via graph-based manifold ranking," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3166-3173, 2013.

[25] Li Y, Hou X, Koch C, et al., "The secrets of salient object segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 280-287, 2014.

[26] Li G, Yu Y, "Visual saliency based on multiscale deep features," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5455-5463, 2015.

[27] Fan D P, Gong C, Cao Y, et al., "Enhanced-alignment measure for binary foreground map evaluation," arXiv preprint arXiv:1805.10421, 2018.

[28] Xie C, Xia C, Ma M, et al., "Complementary trilateral decoder for fast and accurate salient object detection," *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4967-4975, 2021.

[29] Wu Z, Su L, Huang Q, "Cascaded partial decoder for fast and accurate salient object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3907-3916, 2019.

[30] Zhao S, Zhao Y, Li J, "Is depth really necessary for salient object detection," *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1745-1754, 2020.

[31] Wei J, Wang S, Huang Q, "F3 net: Fusion, feedback and focus for salient object detection," *Proceedings of the AAAI Conference on Artifificial Intelligence*, pp. 12321-12328, 2020.

[32] Wu Z, Su L, Huang Q, "Stacked cross refifinement network for edge-aware salient object detection," *Proceedings of the IEEE International Conference on Computer Vision*, pp.7264-7273, 2019.

[33] Ma M, Xia C, Li J, "Pyramidal feature shrinking for salient object detection," *Proceedings of the AAAI Conference on Artifificial Intelligence*, pp. 2311-2318, 2021.