scientific reports

Check for updates

OPEN

Salient object detection with non-local feature enhancement and edge reconstruction

Tao Xu[1]✉, Jingyao Jiang[2], Lei Cai[1], Haojie Chai[1] & Hanjun Ma[3]

The salient object detection task based on deep learning has made significant advances. However, the existing methods struggle to capture long-range dependencies and edge information in complex images, which hinders precise prediction of salient objects. To this end, we propose a salient object detection method with non-local feature enhancement and edge reconstruction. Firstly, we adopt self-attention mechanisms to capture long-range dependencies. The non-local feature enhancement module uses non-local operation and graph convolution to model and reason the region-wise relations, which enables to capture high-order semantic information. Secondly, we design an edge reconstruction module to capture essential edge information. It aggregates various image details from different branches to better capture and enhance edge information, thereby generating saliency maps with more exact edges. Extensive experiments on six widely used benchmarks show that the proposed method achieves competitive results, with an average of Structure-Measure and Enhanced-alignment Measure values of 0.890 and 0.931, respectively.

Salient object detection[1] aims to detect the most visually noticeable areas of an image. It serves as a crucial preprocessing step with applications in computer vision, such as image enhancement[2], object recognition[3], event detection[4], video object segmentation[5], and semantic segmentation[6]. Moreover, research in salient object detection has the potential to promote agricultural applications, such as livestock bone localization[7] and jujube crack detection[8]. Traditional salient object detection methods exhibit fast detection speeds, primarily relying on low-level features such as contrast[9] and background priors[10]. Recently, a variety of deep learning-based architectures have made remarkable progress in computer vision tasks[11–15]. Specifically, convolutional neural networks (CNNs) and Transformers have advanced capabilities in extracting semantic features, enabling more accurate detection in complex scenarios. The accuracy and robustness of salient object detection have been greatly improved[16–18]. However, due to complex object shapes or cluttered backgrounds in images, effectively capturing long-range dependencies and leveraging edge information is challenging, which can result in suboptimal performance.

To address this challenge, some methods use contextual modeling to improve performance. Liu et al.[16] use attention mechanisms to learn relevance among pixels, selectively aggregating context for each pixel. It enhances the saliency reasoning by integrating global and local context. Siris et al.[17] use scene contexts to detect salient objects. It learns detailed semantic information from a scene by segmenting things and stuff. Qin et al.[19] propose a cascade model utilizing ReSidual U-blocks to capture contextual information across various scales. Xie et al.[20] design a two-branch structure for different image resolutions to learn continuous semantics and rich details. Some methods combine multi-scale features to explore more details. The strategy leverages multi-level features to predict saliency maps, with high-level features effectively locating salient objects and low-level features accurately detecting details. Pang et al.[21] design two branches of different resolutions for flexible multi-scale feature fusion via interactive learning. Wu et al.[22] learn features through an extreme downsampling strategy, designing a scale-correlated pyramid convolution to recover details by extracting features by fusing multi-level features. Yao et al.[23] capture saliency cues from each feature layer, extracting and integrating key features to precisely localize salient objects. Zhou et al.[24] utilize multiple U-shaped branches at various scales to extract comprehensive multi-scale feature extraction. Despite these advancements, existing methods often struggle to effectively capture long-range dependencies and adequate edge information, particularly in complex scenes where salient objects may be occluded or obscured by clutter.

To this end, we propose a method with non-local feature enhancement and edge reconstruction strategies for salient object detection. On the one hand, to capture long-range dependencies, we adopt a non-local feature

[1]School of Artificial Intelligence, Henan Institute of Science and Technology, Xinxiang 453003, China. [2]School of Mechanical and Electrical Engineering, Henan Institute of Science and Technology, Xinxiang 453003, China. [3]School of Food Science, Henan Institute of Science and Technology, Xinxiang 453003, China. ✉email: xutao1206@qq.com

enhancement module. First, it applies channel enhancement operation on tokens to suppress noises. Then, non-local operation[25] is performed on these adjacent tokens to capture long-range dependencies. Finally, a graph convolutional network is imported to capture high-order semantic relations between regions, improving comprehension of complex relationships. On the other hand, we introduce an edge reconstruction module to generate more exact edges. The module combines various local features from different branches to reconstruct boundaries of salient objects and finally restore more complete boundaries. This helps the model learn salient object contour information via edge supervision[26]. This task leverages complementarity between salient edge and object information to generate saliency maps with precise object boundaries.

The overall architecture of our method is shown in Fig. 1. Specifically, the input image is split into patches and fed to the T2T-ViT backbone for feature extraction and global context capture. Then, the non-local feature enhancement module is adopted to exchange information between adjacent tokens to learn region-wise relations. Finally, we employ the edge reconstruction module to fuse various local features from different branches to help the salient object detection task progressively recover accurate salient object boundaries. In summary, our contributions comprise:

- We propose a non-local feature enhancement module to further capture long-range dependencies. This module employs a channel enhancement operation to suppress noise by interacting with adjacent tokens. Furthermore, we import graph convolution to capture high-order semantic relations between regions.
- We design an edge reconstruction module to capture adequate edge details, generating precise saliency maps with accurate object boundaries. This module aggregates diverse features from different branches to improve detail and global learning, thus recovering object details and edges.
- The proposed method integrates non-local feature enhancement and edge reconstruction modules, which work synergistically to further extract global and local features, enabling more accurate and integral salient object detection in complex scenes. Both quantitative and qualitative experiments demonstrate the effectiveness of our proposed method.

Related work

Deep learning methods in salient object detection

Recent advancements in salient object detection methods have leveraged both CNNs and Transformers to enhance feature extraction and improve model performance. Qin et al.[19] proposed a two-level nested U-structure with ReSidual U-blocks to capture contextual information across multiple scales. Liu et al.[16] incorporated attention mechanisms to learn pixel relevance and fuse global and local context for better performance. Pang et al.[21] aggregated multi-scale features via interactive learning, using a consistency-enhanced loss to address pixel imbalance between salient and background regions. Wu et al.[22] used extreme downsampling techniques to explore high-level features for salient object localization, progressively recovering multi-level features in the decoder. Xie et al.[20] designed a two-branch structure combining Transformer and convolutional neural network backbones to capture continuous semantics and rich details from different image resolutions, facilitating information transfer between the branches to mitigate common defects in both architectures. Zhao et al.[27] utilized the complementarity of the edge and object to improve accuracy. Tang et al.[28] divided the salient object detection task into a pixel-wise classification task and a refinement task, assigning each task to different networks. The classification network utilizes contour information to distinguish between foreground, background, and

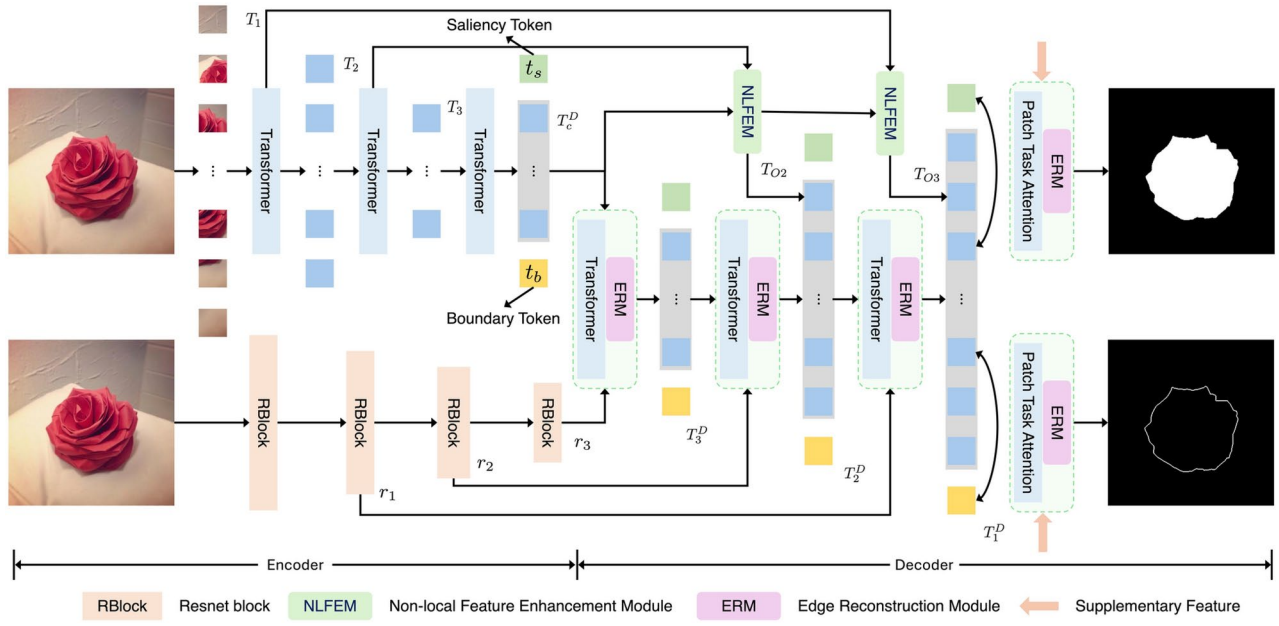

Fig. 1. Overall architecture of the proposed model.

uncertain regions. The refinement network uses the above information as input to get saliency maps with exact boundaries. In our method, edge reconstruction plays a central role, where we integrate features extracted from both convolutional neural network and Transformer branches. The convolutional neural network backbone features are utilized as complementary information to improve the reconstruction of object boundaries. Simultaneously, edge supervision is employed to refine feature precision, further enhancing the accuracy of edge detection.

Attention mechanisms in salient object detection

Attention mechanisms simulates human cognitive processes by selectively focusing on important information and assigning different weights to input data. For example, Liu et al.[16] employed attention mechanisms to learn pixel correlations, selectively aggregating contextual information for each pixel. This approach improves saliency inference by integrating both global and local context. Wang et al.[29] introduced a pyramid attention structure, stacking multiple attention layers to handle multi-scale saliency features. Self-attention mechanisms enable models to focus on different positions within the same sequence, accounting for relationships between various parts. Xie et al.[20] proposed a dual-branch structure for high-resolution detection, using both Transformer and convolutional network to process images at different resolutions, capturing continuous semantics and detailed information. Liu et al.[26] utilized a Transformer architecture with the T2T-ViT backbone to propagate global context and applied the reverse tokens-to-token module to upsample patch tokens for gradual resolution recovery. Chen et al.[18] integrated features from various layers and utilized global contextual information at different stages to prevent the dilution of high-level features. The attention mechanisms and the non-local operation both capture long-range dependencies across spatial locations. In particular, it can directly compute the relationship between any two positions in an image, regardless of their spatial distance. To further acquire contextual information and local details, a non-local feature enhancement module is adopted in our network. This module interacts between adjacent tokens to aggregate nearby features while exploring high-order semantic relations between regions, leading to more robust contextual understanding and improved saliency detection.

Method

In this paper, we propose a salient object detection method with non-local feature enhancement module (NLFEM) and edge reconstruction module (ERM), aiming to predict finer details and more complete salient objects. In this section, we first describe the overall architecture of the network, followed by detailed explanations of the Non-Local Feature Enhancement Module, the Edge Reconstruction Module, the decoder, and the loss function.

Overall architecture

Like Liu et al.[26], we use T2T-ViT[30] as the backbone and adopt an encoder–decoder architecture, as shown in Fig. 1. First, the encoder generates multi-level tokens utilizing self-attention mechanisms, which help capture long-range dependencies. Subsequently, the non-local feature enhancement module is adopted to enhance region-wise relations. It explores high-order semantic relations via non-local operation and graph convolution. Finally, we use features extracted from different branches and supplementary features derived through the processing of I to learn useful edge information and conduct edge reconstruction. In the decoder, features are gradually upsampled to the full resolution.

Specifically, in the encoder, $I \in \mathbb{R}^{W \times H \times C}$ is first divided into overlapping patches via a soft split step. The resulting token sequences from the backbone have different shapes: $T_1 \in \mathbb{R}^{\frac{H \times W}{4 \times 4} \times c}$, $T_2 \in \mathbb{R}^{\frac{H \times W}{8 \times 8} \times c}$, and $T_3 \in \mathbb{R}^{\frac{H \times W}{16 \times 16} \times c}$. Besides, T_c^D is transformed from the encoder space to the decoder space to get T_3^D using transformer layers and the depth is set to 4. These multi-level token sequences are processed by the non-local feature enhancement module, where adjacent token sequences interact to produce enhanced feature tokens: T_{O1}, T_{O2} and T_{O3}. In the decoder, a multi-task framework is employed to perform both saliency and boundary prediction tasks. The token sequence length of T_3^D is progressively restored to its original resolution using the RT2T module. The final saliency and boundary maps are predicted from T_1^D. Additionally, the features extracted from the ResBlock are denoted as r_1, r_2, r_3. The edge reconstruction module is incorporated to capture additional details from these supplementary features. By integrating this module at the end of each decoder, the edge prediction task is enhanced, leading to more accurate predictions.

Non-local feature enhancement module

Transformers have demonstrated impressive performance in capturing long-range dependencies, yet they are limited in learning high-order semantic information between regions. A cluttered background or complex salient object shape leads to similar visual features among salient objects, noise objects, and background, which increases the difficulty of distinguishing them due to subtle clues that cannot be identified.

Inspired by methods[25,31,32], a non-local feature enhancement module is introduced to reason high-order semantic relations between regions, as shown in Fig. 2. A simple channel enhancement operation is first adopted for adjacent token sequences to uncover relations among different channels. Then, a non-local operation is adopted for adjacent tokens for the aggregation of adjacent salient clues. Finally, a graph convolution network is employed to learn high-order semantic relations between regions, exploring subtle discriminative features.

Unlike previous methods focused on spatial dependencies, we introduce a simple yet effective channel enhancement (CE) operation for adjacent token sequences to capture subtle inter-channel relationships often missed by traditional self-attention mechanisms. This strengthens cross-channel interactions, enhancing the model's ability to differentiate between salient objects and background noise. Specifically, the adjacent tokens T_i and T_{i+1} are first integrated into a single token T_c using CE[33]. T_{i+1} is adjusted to the same resolution as

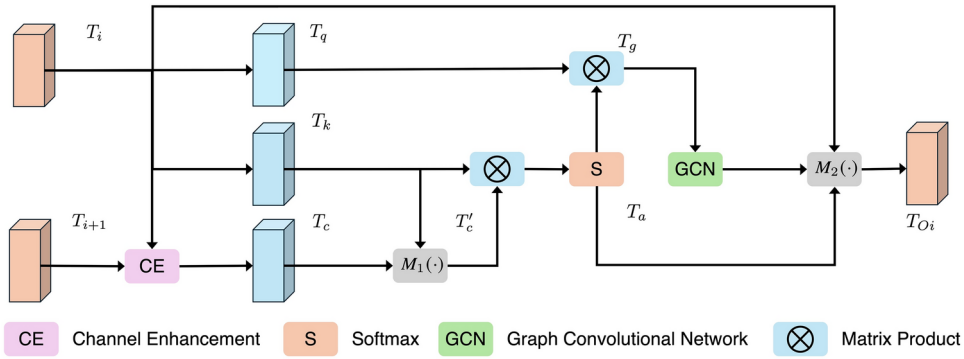

Fig. 2. Non-local feature enhancement module.

T_i using an upsampling operation. Then, the fusion map T_{fuse} is generated by concatenating these tokens $T_{fuse} = \text{Concat}\ [T_i, T_{i+1}, T_i]$. Next, T_{fuse} was applied with l_2 normalization to filter out background noise:

$$T_c = T_{fuse} \odot\ F_{Bottle}(ReLU(LN(F_{Bottle}(T_{fuse})))), \tag{1}$$

where layer normalization is introduced to mitigate the increase in optimization difficulty correlated with channel transformation, F_{Bottle} represents a parameter-efficient bottleneck function, which is employed to reduce the number of parameters while maintaining feature transformation efficiency.

For T_i, two linear projection functions are applied to reduce dimension, which can be denoted as $T_q = \omega_q\,(T_i)$, $T_k = \omega_k\,(T_i)$. The interaction token T'_c, obtained by connecting T_c and T_k, is utilized to interact with the features of T_i for feature enhancement. This can be obtained by the following formula:

$$T'_c = M_1(T_k, T_c) = P(T_k \odot softmax(\omega_c(T_c))), \tag{2}$$

where ω_c is a linear projection function. P means the adaptive pooling operation, which enhances efficiency. To simplify, we use M_1 to represent the overall process. M_1 facilitates efficient pooling of spatial clues, using the adaptive pooling function P to aggregate information, thus enabling the model to better capture fine-grained, salient details. Then, we explore correlations via matrix product between T'_c and T_k, generating an attention map T_a by *softmax*. After that, T_q is projected into the graph domain by T_a via matrix multiplication T_g. A single-layer graph convolutional network is adopted for high-order semantic relationship learning. Specifically, the vertex features T_g are passed through the first-order approximation of spectral graph convolution (GCN) to propagate information across vertices and capture global token representations, i.e., $\hat{T}_g = GCN(T_g) = \text{ReLU}((I - A)T_g w_g)$, where A represents the adjacency matrix defining the graph's structure, and $w_g \in \mathbb{R}^{16\times16}$ is the learnable weight matrix of the GCN. This operation helps in learning high-level semantic relationships among regions. To restore token sequences to original feature dimensions, a deserialization operation is adopted.

$$T_{Oi} = M_2(\hat{T}_g, T_a, T_i) = N(\hat{T}_g \otimes T_a^{\top} + T_i), \tag{3}$$

where $\otimes$ denotes matrix product, N represents the deserialization operation that restores the token's dimensionality. To simplify, we use M_2 to represent the overall process above. It plays a critical role in combining the enhanced global features (from the GCN) with the original token information. By doing so, it ensures that the restored token representations carry both local and global context, enabling more robust feature refinement and enhancing the model's overall understanding of complex scenes.

Decoder

To improve salient object detection, an auxiliary boundary detection task is designed. A multi-task decoder with task-related tokens is adopted to perform both tasks. The patch-task-attention is used to predict saliency maps and boundary maps.

The length of T_3 is challenging to predict a high-quality saliency map from the relatively small length. Therefore, in the decoder, we gradually upscale patch tokens using reverse T2T (RT2T) operations for upsampling. salient object detection methods improve the performance through multi-level feature fusion. Inspired by this, we fuse multi-level tokens in the decoder. It can be represented as:

$$T_i^D = MLP(MSA(Concat(RT2T(T_{i+1}^D), T_{Oi}))), \tag{4}$$

where *MLP* denotes multi-layer perceptron, *MSA* denotes multi-head self-attention, and *Concat* represents concatenation operation.

Inspired by transformer-based methods, we predict the saliency map by adding task-related tokens to the token patch sequences. Many existing salient object detection methods widely adopt boundary detection to improve performance, which inspired us to introduce a boundary detection task. Accordingly, we introduce saliency tokens t_s and boundary tokens t_b, which are introduced as learnable parameters. They have the same length as the feature dimensions of the patch tokens they interact with, ensuring that they can effectively integrate with the feature representations. Specifically, both tokens have a length corresponding to the embedding dimension defined in the model, enabling seamless interaction during the transformer processing. We append saliency tokens t_s and boundary tokens t_b to the patch token sequence T_i^D, and then adopt transformer operation to learn task-related embeddings via interacting with patch tokens. Then we update saliency tokens t_s and boundary tokens t_b to obtain T_{i-1}^D.

The final decoder patch tokens T_1^D and task-related tokens are adopted patch-task-attention for saliency and boundary prediction. For saliency prediction, the T_1^D is embedded into queries Q_s^D, saliency tokens t_s is embedded into a key K_s and a value V_s. The same approach can be applied to the boundary detection task. Then, Task-related patch tokens can be obtained by patch-task-attention:

$$\begin{aligned} T_s^D &= sigmoid\left(Q_s^D K_s^\top / \sqrt{d}\right) V_s + T_1^D, \\ T_b^D &= sigmoid\left(Q_b^D K_b^\top / \sqrt{d}\right) V_b + T_1^D, \end{aligned} \tag{5}$$

where d is the length of V_s and K_s, they have similar lengths. Here, the sigmoid activation is used to calculate attention.

Furthermore, ERM is introduced at the end of each decoder to extract richer edge information. It processes task-related tokens T_D^s and T_D^b to generate the final saliency map and boundary map, ensuring precise edge delineation through saliency supervision and edge supervision. This approach significantly enhances the quality of the output maps.

Edge reconstruction module

Some methods guide the network to excavate edge clues, such as He et al.[34], the paper uses an edge reconstruction module to decode features to generate more exact edges. Some methods introduce edge priors to improve performance. This paper uses supplementary features to complement features from T2T-ViT, thereby addressing the shortcomings in edge feature extraction. This ensures that the edge reconstruction module can adequately learn and leverage edge features. The edge reconstruction module is shown in Fig. 3.

Given features t_i from the decoder and supplementary features r_i extracted from ResBlock or the input image *I*, the ERM is adopted to capture edge information by aggregating these features. The t_i features provide high-level context, while r_i contributes detailed edge information, enhancing edge reconstruction accuracy. For r_i, it is first resized to match t_i for more efficient aggregation. Then, they are added together to further enhance features extracted by Conv-ReLU-Conv to obtain F_1. We use F to denote the Conv-ReLU-Conv framework. For the sum of t_i and r_i, a similar operation F is applied to obtain F_2. To ensure the flexibility of the edge reconstruction module, a weighted gate mechanism g_w is introduced, and a learnable coefficient is obtained by $g_w = S(\sigma_g + \mu_g)$, where σ_g and μ_g are learnable parameters in g_w. Then specific g_w is used to connect different features:

$$\begin{aligned} e_t &= g_{w1} F_1 + (1 - g_{w1}) F_2, \\ F_4 &= g_{w2} F_1 + (1 - g_{w2}) F_3, \end{aligned} \tag{6}$$

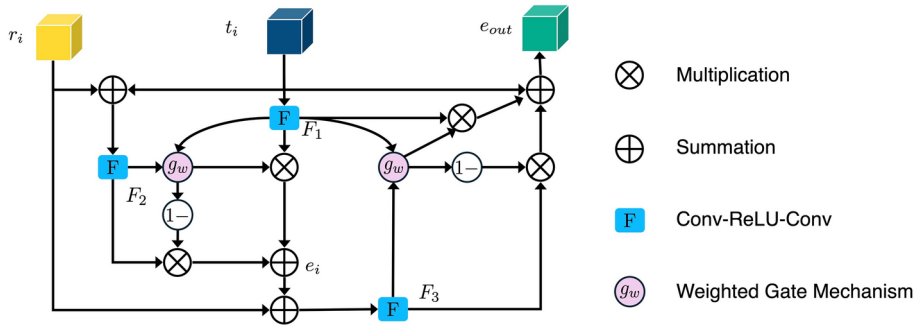

Fig. 3. Edge reconstruction module.

for the sum of e_t and r_i, F is adopted to obtain F_3. Then, similarly, another learnable coefficient is used to learn features flexibly. Here, we nested an ordinary differential equation solver, i.e., a second-order Runge–Kutta[35], to provide more accurate numerical solutions, which can better accommodate the fine-grained property of edges. Next, F_4 and t_i are added to obtain the reconstructed features: $e_{out} = F_4 + t_i$.

Finally, a linear layer is used on e_{out} to generate predictions at different resolutions. Furthermore, we adopt supplementary features through a direct methodology for edge reconstruction. The input image is passed through an ordinary differential equation solver and then aligned with the features obtained by the final decoder. This strategy provides more detail in the final prediction, generating more exact edges.

Loss function

A multi-level hybrid loss L_s is adopted as supervision to learn more details from multi-level features, which combines pixel-level loss $L_{s_{bce}}$, regional-level loss $L_{s_{reg}}$ and object-level loss $L_{s_{obj}}$:

$$L_s = \beta_1 L_{s_{bce}} + \beta_2 L_{s_{reg}} + \beta_3 L_{s_{obj}}, \tag{7}$$

where we set $\beta_1 = \beta_2 = 0.4$, $\beta_3 = 0.2$, following MENet[42]. β_1, β_2 and β_3 represent the weights for the $L_{s_{bce}}$, $L_{s_{reg}}$, and $L_{s_{obj}}$, respectively. By adjusting weight allocation, the model can focus on either local details or global performance, influencing the balance between detail recovery and overall object detection, and affecting final performance. $L_{s_{bce}}$ is a BCE loss which can denoted as:

$$L_{s_{bce}} = -\sum (GlogS + (1-G)log(1-S)), \tag{8}$$

where *G* denote the ground truth (GT), *S* denotes the predicted saliency map. We divide *G* and *S* into four equal sub-regions. Subsequently, we calculate the regional-level loss for each sub-region $L_{s_{Reg}}$ by combining *SSIM* and *IoU*:

$$L_{s_{Reg}} = 1 - \sum_{i=1}^{4} \omega_i (\theta_1 SSIM_i + \theta_2 IoU_i), \tag{9}$$

where $\theta_1 = \theta_2 = 0.5$, following MENet[42]. θ_1 and θ_2 represent the weights for *SSIM* and *IoU*, respectively. By assigning equal weights, the model effectively captures both structural relationships and area overlap between predicted and ground truth maps. ω_i is the ratio of predicted foreground to corresponding ground truth foreground in each region S_i and $G_i (i \in [1, 4])$. The $SSIM_i$ calculates the luminance, contrast, and structure comparison. IoU_i measures overlap of regions between S_i and G_i. The object-level loss calculates the foreground distribution, mainly considering the foregrounds of *S* (i.e., S_o) and *G*. $L_{s_{obj}}$ is defined as:

$$L_{s_{obj}} = 1 - \frac{2\mu_{S_o}}{\mu_{S_o}^2 + 1 + 2\lambda\sigma_{S_o}}, \tag{10}$$

where μ_{S_o} and σ_{S_o} denote the mean and the standard deviation of S_o, λ represents the weight.

Experiments

Datasets

We use DUTS-TR[38] dataset for training, while six commonly used benchmarks: SOD[39], DUT-OMRON[40], PASCAL-S[41], HKU-IS[36], ECSSD[37], and DUTS-TE[38] serve as test datasets to evaluate models. All datasets have pixel-level annotations. SOD[39] (300 images) contains salient objects in natural scenes. Some images have multiple salient objects. DUT-OMRON[40] (5168 images) has multiple salient objects or complex backgrounds in some images. PASCAL-S[41] (850 images) has multiple salient objects and multiple salient values. HKU-IS[36] (4447 images) contains multiple salient objects, with at least one touching the image boundary. They have low contrast. ECSSD[37] (1000 images) has complex structures, multiple categories objects, complex diverse backgrounds and one or more salient objects, possibly transparent. DUTS[38] comprise DUTS-TR[38] (10,553 training images) and DUTS-TE[38] (5019 testing images). The dataset boasts a diverse array of images showcasing various scenes, objects, and backgrounds, posing a challenge to computer vision models.

Evaluation metrics

We select 6 evaluation metrics: Precision-Recall (PR) curve, F-measure[43], MAE[44], weighted F-measure[45], S-measure[46], E-measure[47]. F-measure[43] comprehensively considers Precision and Recall. It can be denoted as:

$$F_\beta = \frac{(1+\beta^2)\text{Precision} \times \text{Recall}}{\beta^2 \text{Precision} + \text{Recall}}, \tag{11}$$

in the salient object detection task, precision and recall should be considered comprehensively, β^2 value is set to 0.3. MAE can evaluate closeness between *S* and *G*. It can be denoted as:

$$\mathrm{MAE} = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} |G(i,j) - S(i,j)|, \tag{12}$$

where $G(i,j)$ and $S(i,j)$ are the value at position (i, j) in G and S, respectively. The weighted F-measure[45] can reflect saliency continuity, differences across locations and relative importance of pixels or regions. It is calculated from precision and recall values:

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

The weighted F-measure[45] is an extension of the F-measure[43]. It assigns different weights ω to errors according to location and neighborhood. S-measure[46] evaluates structural similarity between G and S, emphasizing the global structure of objects, and can be computed as:

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

where α is set to 0.5, S_o and S_r denote object-aware and region-aware structural similarity, respectively. E-measure[47] considers both local and global information, considering pixel position and image-level information. It is denoted as:

$$E_\xi = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} \theta(\xi) \tag{15}$$

where ξ and $\theta(\xi)$ represent the alignment matrix and the enhanced alignment matrix, respectively.

Experimental details

In this paper, the boundary ground truth is generated by sober operator to assist edge supervision. Random flipping is used for data enhancement. The image is randomly cropped to 224×224 after resized to 256×256 as input. The pre-trained T2T-ViT-24 model was adopted as our backbone. We set the training steps as 60,000. The Adam is used as optimizer. The learning rate starts at 0.0001 and decreases by a factor of 10 at 1/2 and 3/4 of the total steps, respectively, like VST[26]. We use PyTorch to implement the model with CUDA 11 environment, trained on an NVIDIA A100 80GB PCIe GPU.

Comparison with state-of-the-art methods

We conduct both quantitative and qualitative comparisons of our method with 17 existing methods, highlighting the differences and advantages. These compared methods adopt different backbones. U2Net[19] uses the RSU[19] as its backbone. EDN[22] and ICON[33] use the VGG-16[48] as backbones. BASNet[49] uses the ResNet-34[50] as backbone. PiCANet[16], SCRN[51], LDF[52], ITSD[53], CTDNet[54], RCSB[55], OLER[23], MENet[42], ELSANet[56], EMSNet[24], DC-Net[57] and CANet[58] use the ResNet-50[50] as backbones. The backbone of VST[26] is T2T-ViT-14[30]. For fair comparisons, we ensured that all methods were evaluated using publicly available datasets, with saliency maps obtained either by running the released codes or from those provided by the authors. Notably, our method performs well in terms of two key metrics, with average S_m and E_m^{max} values of 0.890 and 0.931, respectively, highlighting its ability to accurately capture both global and structural information. In particular, our method demonstrates exceptional performance on challenging datasets such as HKU-IS, ECSSD, and DUT-OMRON, as evidenced by the qualitative comparisons. However, it has limitations with insufficient accuracy in detecting salient objects and inadequate separation from the background in certain rows, primarily due to inadequate learning of local features.

Quantitative comparison

The quantitative comparison presents results for five metrics and the comparison of PR curves, as shown in Tables 1, 2 and Fig. 4, respectively. It is evident that our method achieves or approaches the top-2 performance in five evaluation metrics for six datasets. More importantly, our method performs best on S_m and E_m^{max} for HKU-IS, ECSSD, DUTS-TE, DUT-OMRON and PASCAL-S. Compared to the second-best method, MENet[42], our method demonstrates an average improvement of 1.1% in S_m and E_m^{max} across six datasets. Analysis of the PR curves further highlights that our method achieves a strong balance between precision and recall. The red dotted lines representing our method consistently demonstrate superior performance. Compared with the results of other methods that achieve an advantage in a certain metric on certain methods (e.g., ELSANet[56] in terms of MAE), our method achieves better generalization, which ranks first in the majority of cases and second or third in rare cases. Compared to the recent DC-Net[57], which aggregates feature maps with varying semantic information from multiple encoders, our method achieves an average improvement of 3.6% across five metrics on six datasets. E_m^{max} considers both local and global information, S_m focuses on structural similarity, these two metrics comprehensively reflect the model's ability to capture global structural information. The improvement on the above mentioned two metrics demonstrates the effectiveness of our method in handling complex shapes. Additionally, MAE reflects noise intensity, F_β^{max} responds to various deficiencies in saliency maps, and F_β^ω reflects the continuity of saliency and the differences in saliency levels at different locations. Improvements

Method	HKU-IS					ECSSD					DUTS-TE				
	$S_m \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$E_m^{max} \uparrow$	$S_m \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$E_m^{max} \uparrow$	$S_m \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$E_m^{max} \uparrow$
PiCANet	0.904	0.840	0.043	0.919	0.950	0.917	0.867	0.047	0.935	0.952	0.869	0.756	0.051	0.860	0.920
BASNet	0.909	0.889	0.032	0.928	0.952	0.916	0.904	0.037	*0.942*	0.951	0.866	0.803	0.047	0.860	0.903
SCRN	0.916	0.876	0.034	0.934	0.956	0.927	0.899	0.038	0.950	0.037	0.885	0.803	0.040	0.888	0.925
LDF	0.919	0.904	0.027	0.939	0.958	0.924	0.915	0.034	0.950	0.954	0.892	0.845	0.034	0.898	0.930
U2Net	0.916	0.890	0.031	0.935	0.954	0.928	0.910	0.033	0.951	0.957	0.861	0.804	0.044	0.873	0.911
ITSD	0.917	0.894	0.031	0.934	0.960	0.925	0.911	0.035	0.947	0.959	0.885	0.824	0.041	0.883	0.929
CTDNet	0.922	0.909	0.027	0.941	0.961	0.925	0.915	0.032	0.950	0.956	0.893	0.847	0.034	0.897	0.935
VST	*0.928*	0.897	0.030	0.937	*0.968*	*0.932*	0.910	0.034	0.944	*0.964*	0.896	0.828	0.037	0.877	0.939
RCSB	0.919	0.909	0.027	0.938	0.959	0.922	0.916	0.033	0.944	0.955	0.881	0.840	0.034	0.889	0.925
EDN	0.921	0.900	0.029	0.938	0.959	0.928	0.915	0.034	0.948	0.959	0.883	0.822	0.041	0.881	0.922
ICON	0.915	0.895	0.032	0.935	0.957	0.919	0.905	0.036	0.945	0.953	0.878	0.822	0.043	0.883	0.924
OLER	0.920	0.911	0.042	0.940	0.960	0.927	0.924	0.030	0.953	0.959	0.889	0.852	0.0332	0.896	0.934
MENet	0.927	0.917	**0.023**	**0.948**	0.965	0.928	0.920	0.031	*0.955*	0.956	*0.905*	0.870	**0.028**	**0.912**	*0.944*
ELSANet	0.923	0.916	*0.025*	0.935	0.963	0.929	0.926	*0.030*	0.943	0.960	0.893	0.856	0.034	0.882	0.934
EMSNet	0.921	0.912	*0.025*	0.940	0.960	0.926	0.921	0.031	0.948	0.955	0.893	*0.892*	0.031	0.896	0.933
DC-Net	0.924	0.909	0.027	0.942	0.963	0.924	0.913	0.034	0.949	0.953	0.896	0.852	0.035	0.899	0.935
CANet	0.920	**0.939**	0.027	0.941	0.961	0.928	**0.951**	0.032	0.950	0.958	0.895	**0.898**	0.033	0.899	0.937
Ours	**0.934**	*0.920*	**0.023**	**0.948**	**0.972**	**0.939**	*0.931*	**0.026**	**0.957**	**0.969**	**0.907**	0.866	*0.030*	*0.904*	**0.948**

Table 1. Quantitative comparison on three complex datasets: HKU-IS[36], ECSSD[37] and DUTS-TE[38]. For ↑ and ↓, higher and lower scores indicate better results, respectively. E_m^{max} denotes max E-measure, F_β^{max} denotes max F-measure. The best and second-best results are shown in bold and italics, respectively. The symbol '–' indicates the results of the model are unavailable.

in these three metrics validate our model's capability to address visually confusing scenes. The proposed method achieves good performance, especially on three relatively complex datasets: HKU-IS[36], ECSSD[37] and DUTS-TE[38]. However, the performance on the SOD[39], DUT-OMRON[40], and PASCAL-S[41] datasets is slightly lower, particularly in terms of F_β^ω, MAE, and F_β^{max}, suggesting that the model may occasionally misidentify background regions as salient objects.

Qualitative comparison

Figure 5 shows the qualitative comparison with DC-Net[57], CANet[58], OLER[23], ICON[33], EDN[22], VST[26], CTDNet[54], SCRN[51], and BASNet[49] on challenging images from DUTS-TE (rows 1–3), DUT-OMRON (rows 4–6), and HKU-IS (rows 7–9). Each row, from top to bottom, depicts scenes with occlusion, complex structures, low contrast, small objects, blurred boundaries, intricate shapes, cluttered backgrounds, lighting effects, and multiple salient objects, respectively. Our method effectively captures global structural information while preserving fine-grained details, even in the presence of complex structures and challenging backgrounds. For example, the bird obscured by branches (row 1), bees with transparent wings (row 2), fountains surrounded by splashing water (row 6), and transparent glasses near a soccer ball (row 7) illustrate challenging scenarios. These scenes contain details that are easily confusable and structural information that is challenging to perceive. Moreover, our method effectively distinguishes multiple salient objects from the background. Specifically, in rows 3, 8, and 9, it accurately differentiates birds and cups from their backgrounds, successfully detecting all salient objects. Additionally, the method effectively minimizes background interference. This is demonstrated in rows 4 and 5, where the model suppresses irrelevant elements and enhances focus on the salient objects. However, limitations remain in detection accuracy and local feature learning. In scenes with complex backgrounds and intricate shapes, certain boundaries and fine structures sometimes appear blurred, as seen in rows 1 and 2. Furthermore, confusion between the foreground and background is evident in rows 3 and 9.

Ablation study

To validate the effectiveness of different modules, ablation experiments were conducted on NLFEM, ERM, and loss functions we used. The quantitative comparison results of these four methods on HKU-IS[36], DUT-OMRON[40] and DUTS-TE[38] datasets are shown in Table 3. In the table the first row represents the metrics without any module ablation; the second, third and fourth rows, respectively, represent metrics after ablating different modules and the loss functions. For comparison, we replaced the loss function used in this paper with a simple BCE loss to assess the impact of the loss function on model performance. Ablating the ERM results in a performance decline across all metrics. For instance, on the HKU-IS dataset, S_m and F_ω^{max} decrease by 0.007 and 0.010, respectively, while the MAE on the DUTS-TE dataset increases by 0.005. This demonstrates that removing the ERM reduces the accuracy of salient object detection. The ERM plays a crucial role in capturing detailed information, thereby enhancing detection precision. Ablating the NLFEM also led to a slight decline across all metrics. This indicates that directly fusing features from different levels impedes the model's ability to effectively capture long-range dependencies. Ablating the NLFEM also led to a slight decline across all metrics. This indicates that directly fusing features from different levels impedes the model's ability to effectively capture long-range dependencies. Modifying the loss function also impacts performance. While some metrics remain unchanged, others decline. The BCE loss is less effective in balancing local and global feature learning, which is essential for accurate saliency detection, while the multi-level hybrid loss used in this paper is more effective in integrating the contributions of various feature scales and handling background interference. Overall, the combination of modules achieves the best performance. This synergy leverages the complementary strengths of each module, enabling the model to capture both local and global features effectively. The method enhances accuracy in detecting salient objects and effectively differentiating them from the background.

Qualitative comparison on these three datasets is shown in Fig. 6. These datasets feature challenging scenes, including occluded salient objects, separated object structures, complex backgrounds, and highly similar foreground and background, arranged from top to bottom. Ablating the ERM causes the model to struggle more with accurately delineating object edges and preserving fine details, leading to some blurriness (rows 1 and 4) in certain regions. Moreover, the removal of the ERM can result in a loss of key details, causing confusion between the foreground and background (rows 2, 3, and 5). Ablating the NLFEM also hinders the model's ability to capture the overall structure of salient objects and differentiate between the foreground and background. The absence of long-range dependency modeling and relations between regions provided by the NLFEM makes it difficult for the model to capture global information effectively. The qualitative comparison further reveals an increase in background interference after modifying the loss function, highlighting the critical role of the loss function in controlling the model's ability to distinguish between foreground and background. An unbalanced loss function increases the model's susceptibility to confusion and noise, especially in complex scenes with similar foreground and background elements. Qualitative comparison suggests a decline in the model's ability to capture the overall structure of objects, making it more sensitive to noise, particularly in areas with complex boundaries and fine-grained details. From the table and figure, a consistent trend emerges, showing that each module positively contributes to overall performance. Ultimately, the combination of the two modules and the loss function achieves the best performance.

Conclusion

In this work, we propose a method that combines non-local feature enhancement and edge reconstruction strategies in the field of salient object detection. This method effectively captures long-range dependencies and exploits edge information and details of salient objects, thereby enhancing the performance of salient object detection. We investigate leveraging non-local operations to promote information interaction between neighboring tokens for further capturing long-range dependencies. Additionally, we explore learning diverse

Method	SOD					DUT-OMRON					PASCAL-S				
	$S_m \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$E_m^{max} \uparrow$	$S_m \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$E_m^{max} \uparrow$	$S_m \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$E_m^{max} \uparrow$
PiCANet	0.793	0.723	0.109	0.858	0.866	0.832	0.695	0.065	0.803	0.876	0.854	0.780	0.087	0.881	0.901
BASNet	0.772	0.728	0.114	0.851	0.832	0.836	0.751	0.056	0.805	0.871	0.838	0.793	0.076	0.854	0.886
SCRN	0.792	0.734	0.104	0.867	0.863	0.836	0.720	0.056	0.812	0.875	0.869	0.807	0.064	0.882	0.910
LDF	0.800	0.765	0.093	0.873	0.866	0.839	0.752	0.051	0.820	0.869	0.863	0.822	0.060	0.874	0.908
U2Net	0.786	0.748	0.108	0.861	0.857	0.847	0.757	0.054	0.823	0.880	0.844	0.797	0.074	0.859	0.883
ITSD	0.809	0.777	0.095	*0.880*	0.874	0.840	0.750	0.061	0.824	0.880	0.859	0.812	0.071	0.871	0.908
CTDNet	–	–	–	–	–	0.844	0.762	0.052	0.826	0.881	0.863	0.822	0.061	0.878	0.906
VST	**0.854**	0.778	**0.065**	0.866	**0.902**	*0.850*	0.755	0.058	0.800	*0.888*	*0.873*	0.816	0.067	0.850	0.900
RCSB	0.750	0.730	1.536	0.846	0.813	0.835	0.752	**0.045**	0.809	0.866	0.860	0.816	0.058	0.876	0.906
EDN	0.798	0.767	0.252	0.868	0.864	0.838	0.746	0.057	0.805	0.871	0.860	0.815	0.066	0.875	0.903
ICON	0.814	*0.784*	0.089	0.872	0.866	0.833	0.743	0.065	0.817	0.879	0.861	0.820	0.064	0.878	0.911
OLER	–	–	–	–	–	0.845	0.775	0.054	0.826	0.891	0.858	0.827	0.063	0.877	0.906
MENet	0.809	0.777	0.087	0.878	0.864	*0.850*	0.771	**0.045**	*0.834*	0.879	0.872	0.838	**0.054**	*0.890*	*0.915*
ELSANet	–	–	–	–	–	0.846	0.774	0.050	0.794	0.885	0.864	0.836	0.060	0.862	0.910
EMSNet	0.798	0.819	1.962	0.863	0.852	0.838	0.759	0.048	0.807	0.870	0.868	*0.848*	**0.054**	**0.892**	*0.915*
DC-Net	0.797	0.761	0.100	0.862	0.855	0.849	0.772	0.053	0.827	0.883	0.857	0.818	0.067	0.878	0.899
CANet	0.803	0.777	0.244	0.874	0.869	0.847	**0.828**	*0.047*	0.827	0.886	0.859	**0.888**	0.061	0.876	0.910
Ours	*0.824*	**0.801**	*0.082*	**0.886**	*0.881*	**0.860**	*0.789*	0.051	**0.839**	**0.899**	**0.874**	0.837	*0.057*	0.882	**0.919**

Table 2. Quantitative comparison on SOD[39], DUT-OMRON[40] and PASCAL-S[41] datasets. For ↑ and ↓, higher and lower scores indicate better results, respectively. E_m^{max} denotes max E-measure, F_β^{max} denotes max F-measure. The best and second-best results are shown in bold and italics, respectively. The symbol '–' indicates the results of the model are unavailable.

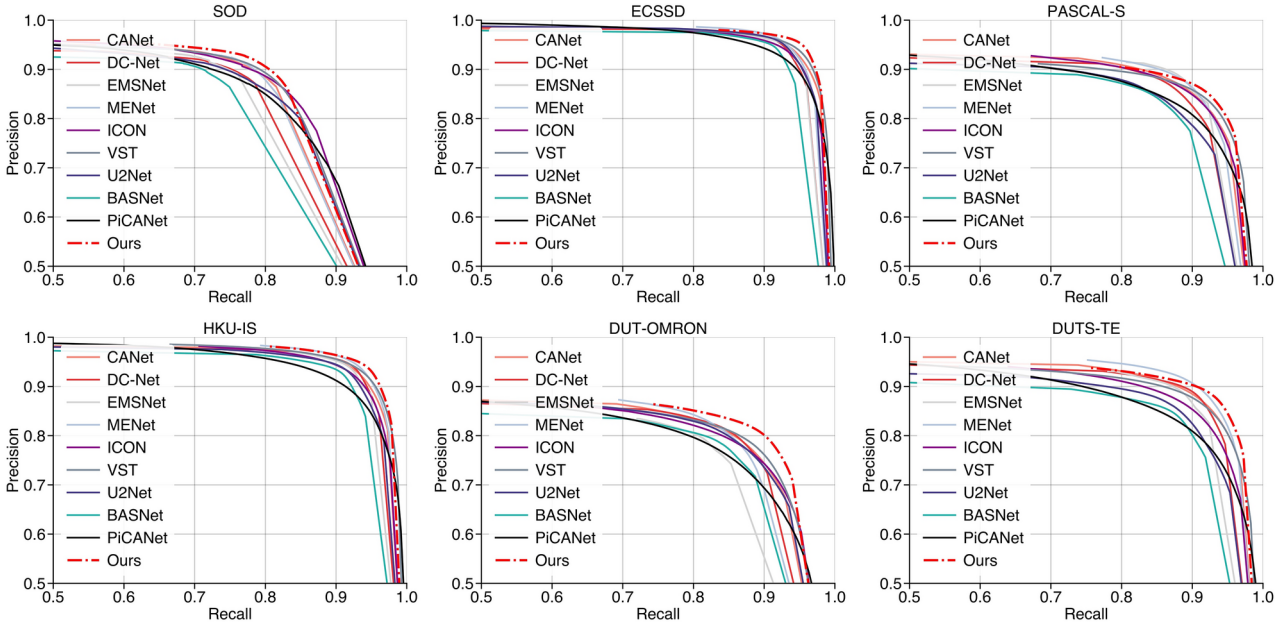

Fig. 4. Precision-recall curves on six benchmarks.

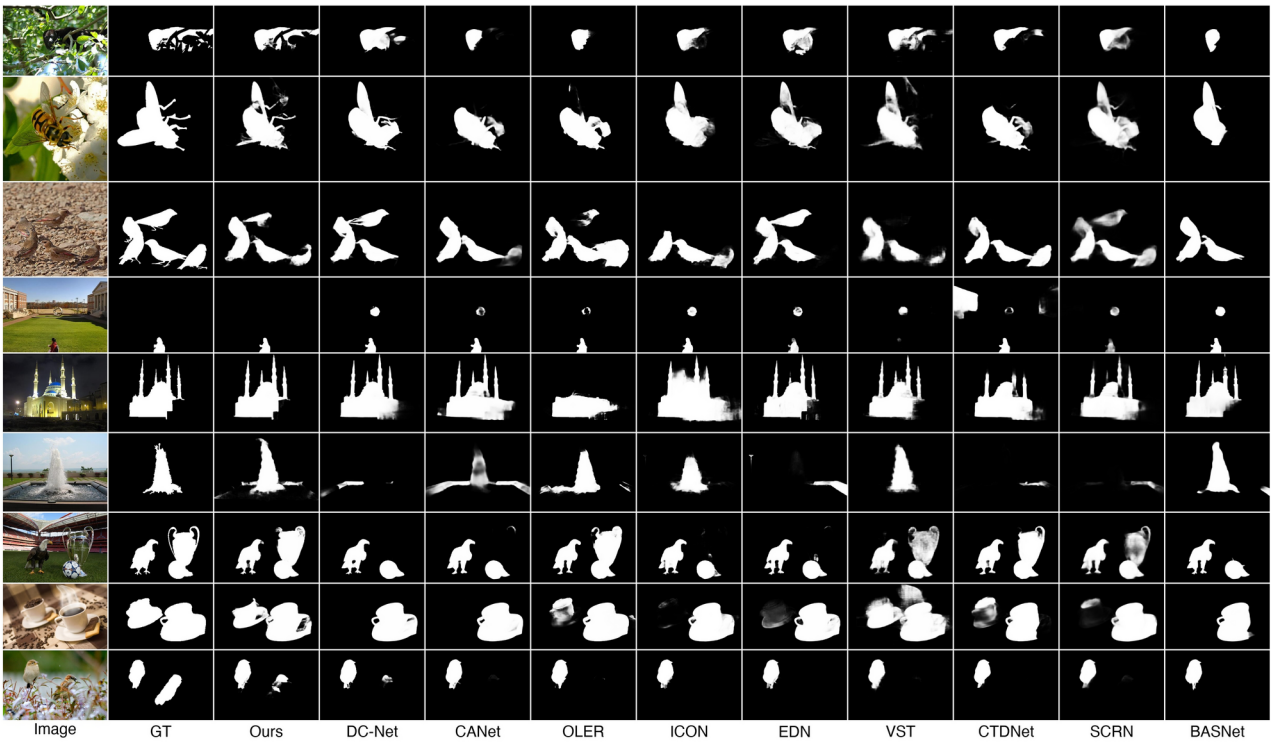

Fig. 5. Qualitative comparison with 9 state-of-the-art methods. Images are sourced from the DUTS-TE[38], DUT-OMRON[40], and HKU-IS[36] datasets.

edge information more effectively from different branches of features to more accurately reconstruct the edges of salient objects. Extensive experiments on six benchmarks demonstrate the outstanding performance of the proposed method on six evaluation metrics. Visually, the prediction maps generated by our method exhibit excellent performance in terms of object and edge integrity. In future work, we will focus on further improving the robustness and speed of salient object detection tasks in specific scenarios to facilitate its application across in industrial fields and real-time systems.

Ablation module	HKU-IS					DUT-OMRON					DUTS-TE				
	$S_m \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$E_m^{max} \uparrow$	$S_m \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$E_m^{max} \uparrow$	$S_m \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$E_m^{max} \uparrow$
Ours	**0.934**	**0.920**	**0.023**	**0.948**	**0.972**	0.860	**0.789**	**0.051**	**0.839**	**0.899**	**0.907**	**0.866**	**0.030**	**0.904**	**0.948**
ERM	0.927	0.910	0.025	0.941	0.969	0.851	0.775	0.057	0.828	0.892	0.894	0.845	0.035	0.886	0.939
NLFEM	0.933	0.917	0.024	0.947	0.971	0.857	0.782	0.055	0.835	0.896	0.905	0.860	0.032	0.901	0.946
Loss	**0.934**	0.908	0.026	**0.948**	0.971	**0.862**	0.777	**0.051**	**0.839**	0.898	**0.907**	0.850	0.033	**0.904**	0.947

Table 3. Comparison on three complex datasets after ablation of different modules. The best results are shown in bold.

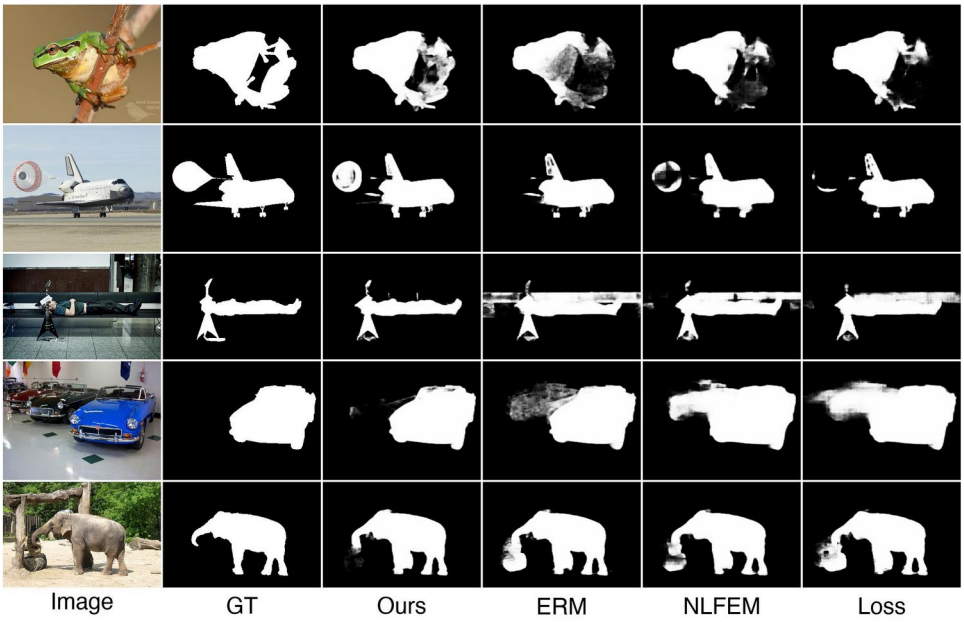

Fig. 6. Qualitative comparison of the ablation study. Images are sourced from the DUTS-TE[38], DUT-OMRON[40], and HKU-IS[36] datasets.

Data availability

The datasets generated and analysed during the current study are available in https://github.com/jiangjingyaocn/ERNet.

Received: 23 April 2024; Accepted: 26 December 2024
Published online: 02 January 2025

References

1. Itti, L., Koch, C. & Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**, 1254–1259 (1998).
2. Miangoleh, S.M.H., Bylinskii, Z., Kee, E., Shechtman, E. & Aksoy, Y. Realistic saliency guided image enhancement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 186–194 (2023).
3. Flores, C. F., Gonzalez-Garcia, A., van de Weijer, J. & Raducanu, B. Saliency for fine-grained object recognition in domains with scarce training data. *Pattern Recogn.* **94**, 62–73 (2019).
4. Gan, C., Wang, N., Yang, Y., Yeung, D.-Y. & Hauptmann, A. G. Devnet: A deep event network for multimedia event detection and evidence recounting. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2568–2577 (2015).
5. Wang, W., Shen, J., Yang, R. & Porikli, F. Saliency-aware video object segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, 20–33 (2018).
6. Chen, T. et al. Saliency guided inter- and intra-class relation constraints for weakly supervised semantic segmentation. *IEEE Trans. Multimed.* **25**, 1727–1737 (2023).
7. Xu, T., Zhao, W., Cai, L., Shi, X. & Wang, X. Lightweight saliency detection method for real-time localization of livestock meat bones. *Sci. Rep.* **13**, 4510. https://doi.org/10.1038/s41598-023-31551-6 (2023).
8. Zheng, Z. et al. AFFU-Net: Attention feature fusion U-Net with hybrid loss for winter jujube crack detection. *Comput. Electron. Agric.* **198**, 107049. https://doi.org/10.1016/j.compag.2022.107049 (2022).
9. Cheng, M.-M., Mitra, N. J., Huang, X., Torr, P. H. S. & Hu, S.-M. Global contrast based salient region detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**, 569–582 (2015).
10. Wei, Y., Wen, F., Zhu, W. & Sun, J. Geodesic saliency using background priors. In *European Conference on Computer Vision*, 29–42 (2012).
11. Liu, D. *et al.* Densernet: Weakly supervised visual localization using multi-scale feature aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 6101–6109 (2021).
12. Wang, W., Han, C., Zhou, T. & Liu, D. Visual recognition with deep nearest centroids. In *International Conference on Learning Representations (ICLR)* (2023).
13. Cui, Y., Yan, L., Cao, Z. & Liu, D. Tf-blender: Temporal feature blender for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8138–8147 (2021).
14. Liu, D., Cui, Y., Tan, W. & Chen, Y. Sg-net: Spatial granularity network for one-stage video instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9816–9825 (2021).
15. Wang, W., Liang, J. & Liu, D. Learning equivariant segmentation with instance-unique querying. *Adv. Neural Inf. Process. Syst.* **35**, 12826–12840 (2022).
16. Liu, N., Han, J. & Yang, M.-H. Picanet: Pixel-wise contextual attention learning for accurate saliency detection. *IEEE Trans. Image Process.* **29**, 6438–6451 (2020).
17. Siris, A., Jiao, J., Tam, G. K., Xie, X. & Lau, R. W. Scene context-aware salient object detection. In *IEEE/CVF International Conference on Computer Vision*, 4136–4146 (2021).
18. Chen, Z., Xu, Q., Cong, R. & Huang, Q. Global context-aware progressive aggregation network for salient object detection. *Proc. AAAI Conf. Artif. Intell.* **34**, 10599–10606 (2020).
19. Qin, X. et al. U2-Net: Going deeper with nested u-structure for salient object detection. *Pattern Recogn.* **106**, 107404 (2020).

20. Xie, C. *et al.* Pyramid grafting network for one-stage high resolution saliency detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11707–11716 (2022).
21. Pang, Y., Zhao, X., Zhang, L. & Lu, H. Multi-scale interactive network for salient object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9410–9419 (2020).
22. Wu, Y.-H., Liu, Y., Zhang, L., Cheng, M.-M. & Ren, B. EDN: Salient object detection via extremely-downsampled network. *IEEE Trans. Image Process.* **31**, 3125–3136 (2022).
23. Yao, Z. & Wang, L. Object localization and edge refinement network for salient object detection. *Expert Syst. Appl.* **213**, 118973 (2023).
24. Zhou, C., Wang, Z., Zhou, Y. & Pan, C. Emsnet: Extremely multi-scale network for salient object detection. *Multimed. Tools Appl.*[SPACE]https://doi.org/10.1007/s11042-024-19503-2 (2024).
25. Wang, X., Girshick, R., Gupta, A. & He, K. Non-local neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7794–7803 (2018).
26. Liu, N., Zhang, N., Wan, K., Shao, L. & Han, J. Visual saliency transformer. In *IEEE/CVF International Conference on Computer Vision*, 4702–4712 (2021).
27. Zhao, J. *et al.* EGNet: Edge guidance network for salient object detection. In *IEEE/CVF International Conference on Computer Vision*, 8778–8787 (2019).
28. Tang, L., Li, B., Zhong, Y., Ding, S. & Song, M. Disentangled high quality salient object detection. In *IEEE/CVF International Conference on Computer Vision*, 3560–3570 (2021).
29. Wang, W., Zhao, S., Shen, J., Hoi, S. C. H. & Borji, A. Salient object detection with pyramid attention and salient edges. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1448–1457 (2019).
30. Yuan, L. *et al.* Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *IEEE/CVF International Conference on Computer Vision*, 538–547 (2021).
31. Te, G., Liu, Y., Hu, W., Shi, H. & Mei, T. Edge-aware graph representation learning and reasoning for face parsing. In *European Conference on Computer Vision*, 258–274 (2020).
32. Huang, Z. *et al.* Feature shrinkage pyramid for camouflaged object detection with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5557–5566 (2023).
33. Zhuge, M. et al. Salient object detection via integrity learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 3738–3752 (2023).
34. He, C. *et al.* Camouflaged object detection with feature decomposition and edge reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22046–22055 (2023).
35. Li, B. *et al.* ODE transformer: An ordinary differential equation-inspired model for sequence generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 8335–8351 (2022).
36. Li, G. & Yu, Y. Visual saliency based on multiscale deep features. In *IEEE Conference on Computer Vision and Pattern Recognition*, 5455–5463 (2015).
37. Yan, Q., Xu, L., Shi, J. & Jia, J. Hierarchical saliency detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1155–1162 (2013).
38. Wang, L. *et al.* Learning to detect salient objects with image-level supervision. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3796–3805 (2017).
39. Movahedi, V. & Elder, J. H. Design and perceptual validation of performance measures for salient object segmentation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 49–56 (2010).
40. Yang, C., Zhang, L., Lu, H., Ruan, X. & Yang, M.-H. Saliency detection via graph-based manifold ranking. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 3166–3173 (2013).
41. Li, Y., Hou, X., Koch, C., Rehg, J. M. & Yuille, A. L. The secrets of salient object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 280–287 (2014).
42. Wang, Y., Wang, R., Fan, X., Wang, T. & He, X. Pixels, regions, and objects: Multiple enhancement for salient object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10031–10040 (2023).
43. Achanta, R., Hemami, S., Estrada, F. & Susstrunk, S. Frequency-tuned salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1597–1604 (2009).
44. Perazzi, F., Krähenbühl, P., Pritch, Y. & Hornung, A. Saliency filters: Contrast based filtering for salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 733–740 (2012).
45. Margolin, R., Zelnik-Manor, L. & Tal, A. How to evaluate foreground maps. In *IEEE Conference on Computer Vision and Pattern Recognition*, 248–255 (2014).
46. Fan, D.-P., Cheng, M.-M., Liu, Y., Li, T. & Borji, A. Structure-measure: A new way to evaluate foreground maps. In *IEEE International Conference on Computer Vision*, 4558–4567 (2017).
47. Fan, D.-P. *et al.* Enhanced-alignment measure for binary foreground map evaluation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, 698–704 (AAAI Press, 2018).
48. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations* (2015).
49. Qin, X. *et al.* Basnet: Boundary-aware salient object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7471–7481 (2019).
50. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 770–778 (2016).
51. Wu, Z., Su, L. & Huang, Q. Stacked cross refinement network for edge-aware salient object detection. In *IEEE/CVF International Conference on Computer Vision*, 7263–7272 (2019).
52. Wei, J. *et al.* Label decoupling framework for salient object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13022–13031 (2020).
53. Zhou, H., Xie, X., Lai, J.-H., Chen, Z. & Yang, L. Interactive two-stream decoder for accurate and fast saliency detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9138–9147 (2020).
54. Zhao, Z., Xia, C., Xie, C. & Li, J. Complementary trilateral decoder for fast and accurate salient object detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, 4967–4975 (2021).
55. Ke, Y. Y. & Tsubono, T. Recursive contour-saliency blending network for accurate salient object detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 1360–1370 (2022).
56. Zhang, L. & Zhang, Q. Salient object detection with edge-guided learning and specific aggregation. *IEEE Trans. Circuits Syst. Video Technol.* **34**, 534–548 (2024).
57. Zhu, J., Qin, X. & Elsaddik, A. Dc-net: Divide-and-conquer for salient object detection. *Pattern Recogn.* **157**, 110903 (2025).
58. Zhu, G., Wang, L. & Tang, J. Learning discriminative context for salient object detection. *Eng. Appl. Artif. Intell.* **131**, 107820 (2024).

Acknowledgements

This work was supported by the Key Research and Development Project in Henan Province [231111220700, 241111110200], the Major Science and Technology Project in Henan Province [221100110500], the National Natural Science Foundation of China [62273132].

Author contributions

All authors reviewed the manuscript. T.X. contributed to research directions and ideas. J.J. conducted the experiments. J.J. reviewed and edited the original document. H.C. and H.M. contributed to the original draft preparation. L.C. provided experimental equipment. All authors have read and agreed to the submitted version of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to T.X.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.