

Improve Salient Object Detection via Adaptive Boundary-Location Feature Fusion

Yuanbin Fu

ybfu@suda.edu.cn

Soochow University

Jiahao Wang

Tianjin University

Article

Keywords: saliency detection, feature fusion, feature augmentation, cross-attention

Posted Date: October 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11172244/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Improve Salient Object Detection via Adaptive Boundary-Location Feature Fusion

Yuanbin Fu^{1*} and Jiahao Wang²

^{1*}School of Future Science and Engineering, Soochow University, No.1
Jiuyong West Road, Suzhou, 215222, China.

²School of Computer Software, Tianjin University, No. 135 Yaguan Road,
Tianjin, 300350, China.

*Corresponding author(s). E-mail(s): ybfu@suda.edu.cn;
Contributing authors: wjhwtt@tju.edu.cn;

Abstract

Salient object detection aims to identify objects in natural images that most catch human attention. A common approach in this field involves integrating diverse features learned from multiple network branches. Intuitively, to enhance accuracy, the fusion process shall adapt to individual inputs with varying patterns. To improve such adaptability, this work develops a boundary-location information fusion framework. Specifically, we decompose the original problem into two simplified sub-tasks: object localization and boundary refinement, each tackled by a dedicated branch. To mitigate the negative impact of misleading attributions in the feature space, we further propose an adaptive feature augmentation technique, which dynamically pinpoints potential attributions causing incorrect predictions, thereby enhancing robustness. Additionally, an adaptive cross-attention module is devised with parameters that adjust dynamically to effectively aggregate boundary and location features. Comprehensive quantitative and qualitative results demonstrate the effectiveness of our design, and its superiority over other state-of-the-art competitors.

Keywords: saliency detection, feature fusion, feature augmentation, cross-attention

1 Introduction

Salient object detection (SOD) aims at segmenting visually prominent objects from non-salient objects and backgrounds, serving as a critical precursor for high-level vision tasks [1, 2]. Recently, with the advent of deep learning, end-to-end deep neural networks that

directly learn mappings from raw images to ground truths have become the standard in the field of SOD [3–6]. Despite substantial progress, most existing studies [7–13] train a single model supervised solely by object-level saliency maps. However, this strategy often results in sub-optimal accuracy [6, 14], as a single, unified model struggles to simultaneously encapsulate the diverse knowledge needed for object localization and boundary refinement, which respectively demand distinct capabilities: global semantic information invariant to intensity changes, and local details sensitive to such variations.

For extracting diverse features, several studies have introduced additional branches [14–19] or auxiliary edge supervisions [20–27]. However, a critical oversight of these works lies in their insufficient adaptability to the varied properties and patterns present in different inputs. In essence, while certain feature types may be beneficial for specific scenarios, they can prove detrimental or misleading in others. For example, orientation information may aid in recognizing objects aligned in the same direction, but it becomes counterproductive when objects are oriented differently. These detrimental features, extracted by different branches, can introduce adverse effects during the fusion process. To address this issue, a pivotal question, *i.e.*, *how to effectively explore and fuse useful/valuable information tailored to the unique requirements of individual images*, needs to be answered.

To meet the demands of distinct images, we tackle the aforementioned question by adaptively exploring valuable information based on different inputs. To the best of our knowledge, existing multi-branch SOD models seldom account for adaptability, which stems from two reasons: 1) **Static Network Parameters**. Fixed and static network parameters manifest a limited scope of knowledge, making them inflexible for dynamically adjusting feature extraction and fusion processes tailored to each input instance; and 2) **Lack of Interpretability**. The deep features learned by neural networks often lack interpretability, hindering the understanding of features associated with different images.

Hence, to enhance the adaptability, we develop a boundary-location fusion framework, integrating boundary and location features from varying inputs. This logical partition strategy simplifies the original SOD task into two manageable sub-tasks: object localization and boundary refinement. Based on this, our framework incorporates two specialized branches: a boundary branch to extract boundary-related features and a location branch to extract location-related features, respectively. To mitigate the adverse effects of disruptive factors that vary across different images, we propose a novel adaptive feature augmentation (AFA) technique, which is applied to both the boundary and location features before fusing them. Additionally, we employ an adaptive cross-attention mechanism to effectively fuse the augmented boundary and location features, harnessing their complementary strengths. Our contributions can therefore be summarized as:

- **Adaptive Feature Augmentation:** We propose an adaptive feature augmentation technique to adaptively address the detrimental attributions leading to incorrect predictions, conditioned on different inputs.
- **Adaptive Cross-Attention:** We introduce an adaptive cross-attention to connect boundary and location features, where queries, keys, and values are adaptively computed based on different images.

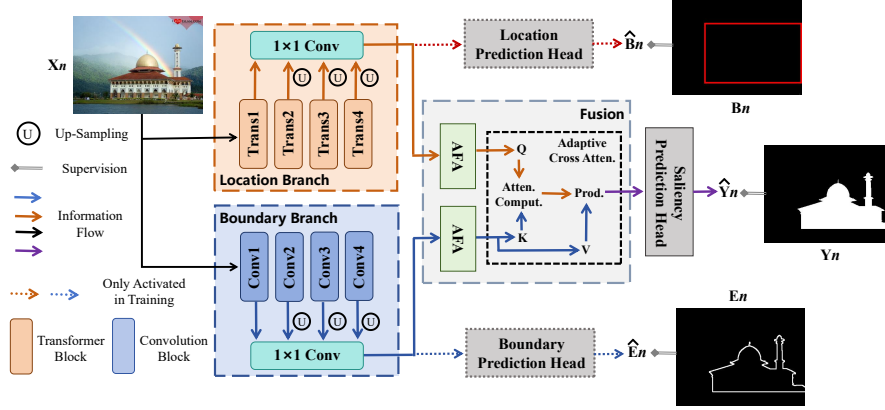


Fig. 1: The overall architecture of our framework, where dotted lines indicate components that are active only during the training phase.

2 Methodology

We denote the training data by $\mathbf{D} := \{(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{B}_n, \mathbf{E}_n), n \in \{1, 2, \dots, N\}\}$, where N is the number of training samples. $\mathbf{X}_n$ is the input image, while $\mathbf{Y}_n$, $\mathbf{B}_n$, and $\mathbf{E}_n$ are the corresponding saliency, location (bounding box), and boundary ground truth, respectively.

2.1 Dual-Branch Network Architecture

In the SOD task, it is imperative to exploit the synergistic advantages of information derived from distinct branches. The logical partition of the complex original problem into simpler sub-problems facilitates the achievement of desirable detection results. Accordingly, as illustrated in Figure 1, this work adopts a dual-branch framework, with each branch dedicated to the extraction of features relevant to boundaries and locations, respectively. In details, we employ a convolution neural network (CNN)-based boundary branch for boundary refinement, and a vision transformer (ViT)-based location branch for object localization. The rationale behind this design is as follows: 1) CNNs are well-suited for extracting local detailed features due to their inherent ability to capture localized information within a convolutional kernel or window; and 2) ViTs excel at modeling global interdependencies among patch tokens, making them effective for exploring global semantic features. We utilize the first four blocks of pre-trained MobileNetV2 [28] and the first four blocks of pre-trained P2T [29] as the backbones of our boundary and location branches, respectively. Both the boundary-related and location-related features are processed by the proposed adaptive feature augmentation (AFA) module before fusion. Subsequently, the boundary and location features are fused through our adaptive cross-attention mechanism. The fused features are then passed through the saliency prediction head, which comprises a series of convolution layers followed by a Sigmoid activation layer to generate the final saliency maps. Note that the boundary and location prediction heads are used solely in training, and will be disabled in testing.

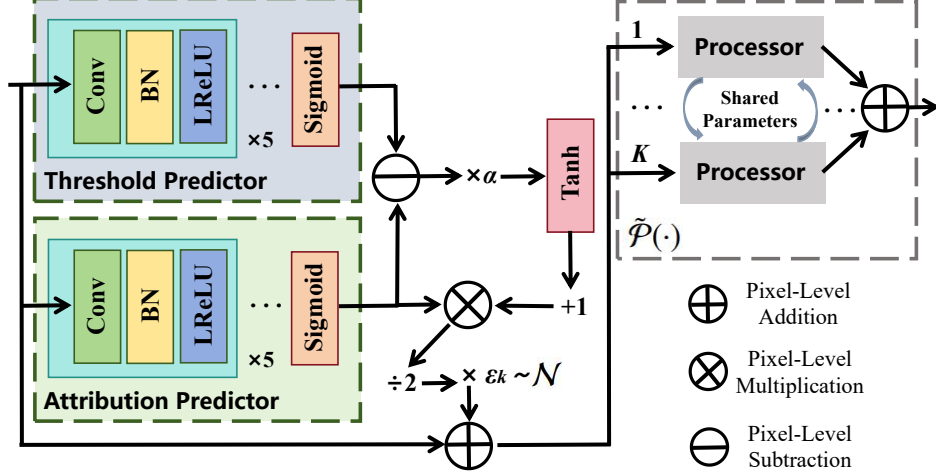


Fig. 2: The proposed adaptive feature augmentation module.

2.2 Adaptive Feature Augmentation

Prior to fusion, we employ an adaptive feature augmentation module, as outlined in Figure 2, to enhance robustness against potential irrelevant/disruptive attributions. This augmentation frees the subsequent fusion module from the adverse effects of misleading features. As demonstrated by previous studies [30, 31], injecting random noises into intermediate features can serve as a form of feature augmentation, helping to prevent overfitting to biased or unreliable information. Building on this concept, we define the feature augmentation function $\mathcal{A}(\mathbf{F}; \mathbf{A}; \varepsilon)$ as:

$$\mathcal{A}(\mathbf{F}; \mathbf{A}; \varepsilon) := \mathbf{F} + \varepsilon \cdot \mathbf{A}, \quad (1)$$

where ε stands for a random scalar sampled from a zero-mean Gaussian distribution $\mathcal{N}$. The attribution matrix $\mathbf{A} \in \mathbb{R}^{H \times W \times D}$ assigns probabilities to each element in the feature map $\mathbf{F} \in \mathbb{R}^{H \times W \times D}$, representing the likelihood that incorrect predictions could be attributed to these elements. H , W , and D mean the height, width, and channel number of features, respectively. Elements with higher values in $\mathbf{A}$ are more likely to correspond to misleading or disruptive attributions, whereas lower values indicate less likelihood of such attributions. However, the unexplainable nature of learned features makes it infeasible to manually assign values to $\mathbf{A}$. Thus, it is essential to proactively learn to identify misleading attributions in the feature space, conditioned on input distributions.

Attribution Estimation. We introduce an attribution predictor to dynamically estimate attribution scores $\mathbf{A}$, as well as a threshold predictor to compute thresholds $\mathbf{T}$ for the subsequent thresholding operation $\mathcal{T}(\cdot)$, both of which consist of five Conv-BatchNorm-LeakyReLU layers followed by a Sigmoid activation. The thresholding operation $\mathcal{T}(\cdot)$ can therefore be written as:

$$\mathcal{T}(\mathbf{A}) := \mathbf{A} \circ (\mathbf{A} > \mathbf{T}), \quad (2)$$

where the comparison operator ($>$) is applied element-wise, ensuring that only the features with higher attribution scores than thresholds are retained for further processing. Rather than

relying on a fixed and unchangeable threshold across all images, we propose to adaptively adjust the thresholds according to each individual input instance, which underpins our usage of a threshold predictor. However, the binary thresholding operation is non-differentiable in nature, making it impractical to be directly optimized through back-propagation.

Hence, in this paper, we investigate an alternative solution $\hat{\mathcal{T}}_\alpha(\cdot)$, which is formulated as:

$$\hat{\mathcal{T}}_\alpha(\mathbf{A}) := \frac{1}{2} \mathbf{A} \circ [\mathbf{1} + \text{Tanh}(\alpha \cdot (\mathbf{A} - \mathbf{T}))], \quad (3)$$

where α is a scalar that controls the steepness of the thresholding function. If α is sufficiently large, $\hat{\mathcal{T}}_\alpha(\cdot)$ can be regarded as a differentiable approximation of $\mathcal{T}(\cdot)$. The following gives the proof:

$$\begin{aligned} \mathcal{T}(\mathbf{A}) &:= \mathbf{A} \circ (\mathbf{A} > \mathbf{T}) = \frac{1}{2} \mathbf{A} \circ [\mathbf{1} + \text{Sign}(\mathbf{A} - \mathbf{T})] \\ &\approx \frac{1}{2} \mathbf{A} \circ [\mathbf{1} + \text{Tanh}(\alpha \cdot (\mathbf{A} - \mathbf{T}))] := \hat{\mathcal{T}}_\alpha(\mathbf{A}). \end{aligned} \quad (4)$$

As α increases, $\text{Tanh}(\alpha \cdot (\mathbf{A} - \mathbf{T}))$ approaches $\text{Sign}(\mathbf{A} - \mathbf{T})$ [32, 33], say $\lim_{\alpha \rightarrow +\infty} \text{Tanh}(\alpha \cdot (\mathbf{A} - \mathbf{T})) = \text{Sign}(\mathbf{A} - \mathbf{T})$.

Feature Augmentation. The reformulated feature augmentation process incorporates the identified misleading/disruptive attributions, $\hat{\mathcal{T}}_\alpha(\mathbf{A})$, to selectively augment the features most likely to introduce bias or interference, which alleviates the potential bias/interference arising from isolated feature points. The updated formulation of Eq. (1) is expressed as:

$$\mathcal{A}(\mathbf{F}; \mathbf{A}; \varepsilon) := \mathbf{F} + \varepsilon \cdot \hat{\mathcal{T}}_\alpha(\mathbf{A}). \quad (5)$$

To substantiate the mathematical soundness of the robustness, we first denote by $\mathcal{P}(\cdot)$ a learnable feature processor (the gray part in Figure 2) applied to the (augmented) features, without loss of generality. After training with multiple groups of augmented features, we can reach:

$$\mathcal{P}(\mathbf{F}) = \mathbb{E}_{\varepsilon \sim \mathcal{N}}[\mathcal{P}(\mathcal{A}(\mathbf{F}; \mathbf{A}; \varepsilon))], \quad (6)$$

where ε is sampled from a zero-mean Gaussian distribution $\mathcal{N}$. We claim that $\mathcal{P}(\cdot)$ possesses the robustness against the augmentation $\mathcal{A}(\mathbf{F}; \mathbf{A}; \varepsilon)$ because:

$$\begin{aligned} \mathcal{P}(\mathcal{A}(\mathbf{F}; \mathbf{A}; \varepsilon_1)) &= \mathbb{E}_{\varepsilon_1, \varepsilon_2 \sim \mathcal{N}}[\mathcal{P}(\mathcal{A}(\mathcal{A}(\mathbf{F}; \mathbf{A}; \varepsilon_1); \mathbf{A}; \varepsilon_2))] \\ &= \mathbb{E}_{\varepsilon_1 + \varepsilon_2 \sim \mathcal{N}}[\mathcal{P}(\mathcal{A}(\mathbf{F}; \mathbf{A}; \varepsilon_1 + \varepsilon_2))] \\ &= \mathcal{P}(\mathbf{F}). \end{aligned} \quad (7)$$

Since both ε_1 and ε_2 are independently sampled from a zero-mean Gaussian distribution, $\varepsilon_1 + \varepsilon_2$ still follows a zero-mean Gaussian, albeit with a different variance.

Due to the impracticability of directly computing the analytic expectation, we approximate Eq. (6) via:

$$\tilde{\mathcal{P}}(\mathbf{F}) := \frac{1}{K} \sum_{k=1}^K \mathcal{P}(\mathcal{A}(\mathbf{F}; \mathbf{A}; \varepsilon_k)), \quad (8)$$

where ε_k is the k -th randomly sampled parameter.

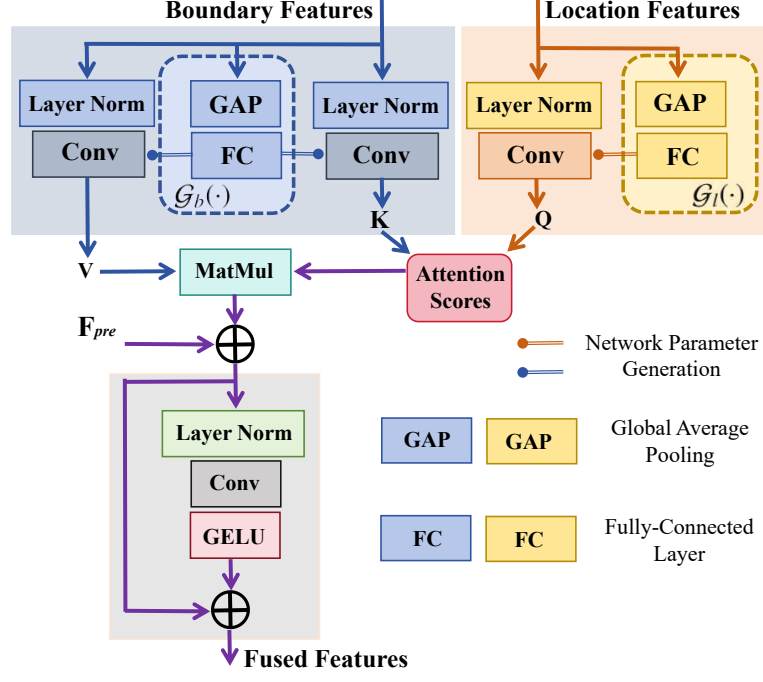


Fig. 3: The proposed adaptive cross-attention module.

2.3 Adaptive Cross-Attention Mechanism

As shown in Figure 3, we propose an adaptive cross-attention mechanism to dynamically explore the complementarity between location and boundary features, which is:

$$\mathbf{F}_{att} := \mathbf{F}_{pre} + \text{Softmax}((\mathbf{F}_l \mathbf{W}_Q)(\mathbf{F}_b \mathbf{W}_K)^T)(\mathbf{F}_b \mathbf{W}_V), \quad (9)$$

where $\mathbf{F}_b$ and $\mathbf{F}_l$ are the boundary and location features augmented by our AFA, respectively. $\mathbf{F}_{pre}$ is obtained by fusing $\mathbf{F}_b$ and $\mathbf{F}_l$ via a 1×1 convolution. $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are linear projection parameters for the query, key, and value, respectively.

Given that different input images present distinct features and patterns, the fusion process should be adjusted dynamically to account for varying input distributions. To realize this, we generate the parameters of the linear projections, *i.e.*, $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$, tailored to the learned boundary and location features. We employ global average pooling (GAP) to aggregate the spatial information of entire feature maps. After GAP, fully-connected layers are utilized to produce the projection parameters. This procedure is described as:

$$\mathbf{W}_Q \leftarrow \mathcal{G}_l(\mathbf{F}_l); \quad [\mathbf{W}_K; \mathbf{W}_V] \leftarrow \mathcal{G}_b(\mathbf{F}_b), \quad (10)$$

where $\mathcal{G}_l(\cdot)$ and $\mathcal{G}_b(\cdot)$ are responsible for generating parameters based on $\mathbf{F}_l$ and $\mathbf{F}_b$, respectively. It is worth mentioning that, the global contextual semantics by the location branch are advantageous in uncovering the relationships among tokens, while local details by the

boundary branch assist in furnishing fine representations containing the local information for individual tokens. Generally, the attention mechanism allows the query to assess long-range dependencies between transformer tokens [34]. In contrast, the key and value encapsulate the local context and the internal representations within individual tokens [35–38]. Hence, we derive the query from the extracted location features, while the key and value from the boundary features.

2.4 Learning Objective

The final learning objective, denoted as $\mathcal{L}_{Final}$, is comprised of four parts, including: saliency regularization $\mathcal{L}_{Sal}$, attribution regularization $\mathcal{L}_{Attr}$, boundary regularization $\mathcal{L}_{Boun}$, and localization regularization $\mathcal{L}_{Local}$, which is formulated as:

$$\mathcal{L}_{Final} := \mathcal{L}_{Sal} + \mathcal{L}_{Attr} + \mathcal{L}_{Boun} + \mathcal{L}_{Local}. \quad (11)$$

Saliency Regularization. We use a binary cross entropy and intersection over union (IOU) loss to measure the distance between predicted saliency maps and ground truths. These two losses are defined as:

$$\mathcal{L}_{CE} := \|\mathbf{Y}_n \circ \text{Log}(\hat{\mathbf{Y}}_n)\|_1 + \|(\mathbf{1} - \mathbf{Y}_n) \circ \text{Log}(\mathbf{1} - \hat{\mathbf{Y}}_n)\|_1, \quad (12)$$

$$\mathcal{L}_{IOU}^S := \left\| \mathbf{1} - \frac{\hat{\mathbf{Y}}_n \circ \mathbf{Y}_n}{\hat{\mathbf{Y}}_n + \mathbf{Y}_n} \right\|_1, \quad (13)$$

where $\circ$ denotes the Hadamard product, and $\|\cdot\|_1$ means the L1 norm. The overall regularization for saliency prediction is:

$$\mathcal{L}_{Sal} := \mathcal{L}_{CE} + \mathcal{L}_{IOU}^S. \quad (14)$$

Attribution Regularization. As previously discussed, to avoid the generation of $\mathbf{A}_n$ and $\mathbf{T}_n$ resulting in trivial values (all-zero or all-one), we impose a regularization term on $\mathbf{A}_n$ and $\mathbf{T}_n$ through penalizing the negative L2 norm as:

$$\mathcal{L}_{Attr} := -\|\mathbf{A}_n \circ (\mathbf{1} - \mathbf{T}_n) \circ \mathbf{T}_n\|_2^2. \quad (15)$$

By Eq. (15), the attribution scores are encouraged to exceed their corresponding thresholds, unless the attributions have minimal effects on erroneous predictions.

Boundary Regularization. We use a binary cross entropy loss and a Dice loss [39] to regularize the boundary prediction $\hat{\mathbf{E}}_n$, ensuring that it closely matches the ground truth $\mathbf{E}_n$. These two losses are defined as:

$$\mathcal{L}_{CE} := \beta_1 \cdot \|\mathbf{E}_n \circ \text{Log}(\hat{\mathbf{E}}_n)\|_1 + \beta_2 \cdot \|(\mathbf{1} - \mathbf{E}_n) \circ \text{Log}(\mathbf{1} - \hat{\mathbf{E}}_n)\|_1, \quad (16)$$

$$\mathcal{L}_{Dice} := \left\| \mathbf{1} - \frac{2 \times \hat{\mathbf{E}}_n \circ \mathbf{E}_n}{|\hat{\mathbf{E}}_n| + |\mathbf{E}_n|} \right\|_1, \quad (17)$$

Table 1: Quantitative comparisons in terms of Max F-measure (F_{max}), and Mean Absolute Error (MAE). Best results are **Bolded**.

Method	Pub.'Year	DUTS		DUT-O		HKU-IS		ECSSD		PASCAL-S	
		$F_{max} \uparrow$	MAE $\downarrow$	$F_{max} \uparrow$	MAE $\downarrow$	$F_{max} \uparrow$	MAE $\downarrow$	$F_{max} \uparrow$	MAE $\downarrow$	$F_{max} \uparrow$	MAE $\downarrow$
GateNet [41]	ECCV'20	.883	.040	.806	.055	.931	.034	.945	.041	.869	.068
LDF [42]	CVPR'20	.896	.033	.819	.051	.939	.027	.950	.033	.874	.059
VST [43]	ICCV'21	.890	.037	.824	.057	.942	.029	.950	.032	.875	.067
SelfReformer [44]	arXiv'22	.916	.027	.836	.041	.947	.024	.957	.027	.894	.050
EDN [45]	TIP'22	.893	.035	.821	.050	.940	.026	.950	.033	.879	.062
BBRF [46]	TIP'23	.916	.027	.843	.044	.953	.022	.959	.026	.891	.054
EVP [47]	CVPR'23	.923	.027	.848	.042	.952	.024	.958	.028	.876	.053
MENet [17]	CVPR'23	.912	.028	.833	.045	.948	.023	.954	.030	.889	.053
CDP [48]	TMM'24	.916	.026	.856	.041	.947	.024	.957	.027	.894	.050
MDSAM [49]	ACM MM'24	.937	.024	.887	.039	.963	.019	.974	.021	.907	.052
DC-Net [50]	PR'25	.930	.022	.857	.039	.956	.021	.967	.020	.899	.051
Ours	-	.939	.021	.890	.038	.966	.016	.978	.018	.911	.047

where $|\cdot|$ is the absolute calculation, and $\|\cdot\|_1$ means the L1 norm. β_1 and β_2 counter imbalanced edge and non-edge pixels [40]. Overall, the regularization for boundary refinement is given by the sum of these two terms:

$$\mathcal{L}_{Boun} := \mathcal{L}_{CE} + \mathcal{L}_{Dice}. \quad (18)$$

Localization Regularization. The location head outputs the coordinates and sizes of the bounding boxes $\hat{\mathbf{B}}_n$ that localize the salient objects. Specifically, $\hat{\mathbf{B}}_n$ is a four-dimension vector, *i.e.*, $\hat{\mathbf{B}}_n := [\hat{x}, \hat{y}, \hat{w}, \hat{h}]$, representing the left-corner coordinate and size of the detected bounding box. Following previous object detection methods, we regularize the detected bounding boxes using a coordinate loss and an Intersection over Union (IOU) loss. These two losses are formulated as:

$$\mathcal{L}_{Coor} := \left\| [\hat{x}, \hat{y}, \sqrt{\hat{w}}, \sqrt{\hat{h}}] - [x, y, \sqrt{w}, \sqrt{h}] \right\|_2^2, \quad (19)$$

$$\mathcal{L}_{IOU} := 1 - \frac{\hat{\mathbf{B}}_n \cap \mathbf{B}_n}{\hat{\mathbf{B}}_n \cup \mathbf{B}_n}, \quad (20)$$

where $\mathbf{B}_n := [x, y, w, h]$ is the ground truth bounding box. Moreover, $\cap$ and $\cup$ represent intersection and union operations, respectively. The combination of the above two terms regularizes object localization as:

$$\mathcal{L}_{Local} := \mathcal{L}_{Coor} + \mathcal{L}_{IOU}. \quad (21)$$

3 Experimental Validation

3.1 Implementation Details

Our model is implemented using the PyTorch library. All experiments are trained on two RTX 3090 GPUs using Adam optimizer [51] with parameters $\beta_1 := 0.9$, $\beta_2 := 0.99$, and weight decay 10^{-9} . The batch size is set to 24, and the poly learning rate scheduler is adopted to

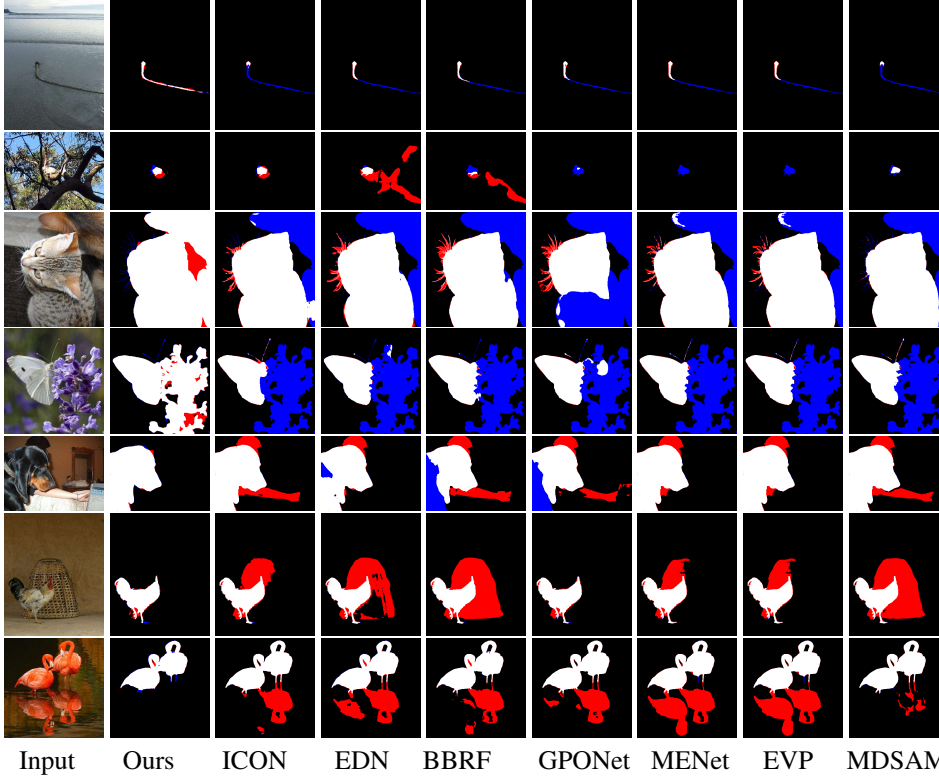


Fig. 4: Visual comparisons between competitors. The red and blue regions represent false-positive and false-negative predictions, respectively.

progressively reduce the learning rate as training proceeds. We adopt MobileNetV2 [28] and P2T [29] as the backbone networks, both of which are initialized with pre-trained parameters on the ImageNet [52]. In the testing phase, the injected ε is fixed to be 0, 0.5, and 1 to ensure deterministic outputs for identical inputs. This setup facilitates consistent evaluation of the model’s performance across different test scenarios.

3.2 Competitors & Datasets & Evaluation Metrics

The competitors for comparison include GateNet [41], LDF [42], VST [43], SelfReformer [44], EDN [45], BBRF [46], EVP [47], MENet [17], CDP [48], MDSAM [49], and DC-Net [50]. Our model is trained on the DUTS-TR dataset [53], and evaluated on DUTS [53], DUT-O [54], HKU-IS [55], ECSSD [56], PASCAL-S [57]. To assess the performance of all SOD models, we employ two commonly used metrics: 1) Max F-measure score (F_{max}) [58]; 2) Mean Absolute Error (MAE).

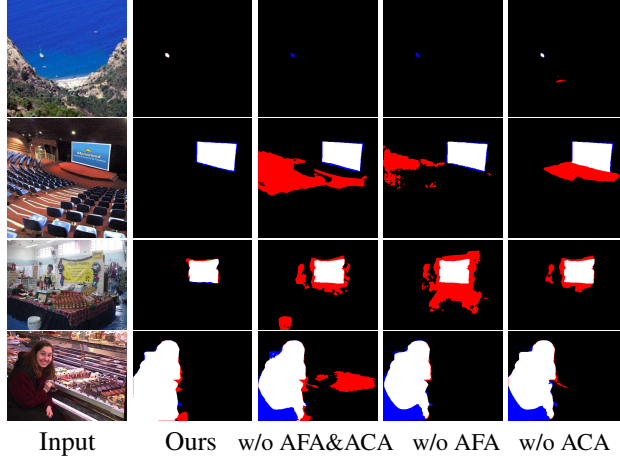


Fig. 5: Visual results of ablation studies. The red and blue regions represent false-positive and false-negative predictions, respectively.

3.3 Comparisons with other methods

We summarize the numerical results in Table 1. It can be seen that, our proposed model significantly outperforms involved competitors on all five datasets in terms of all evaluation metrics. Fundamentally, our high performance across various datasets can be ascribed to the capacity to tailor processing strategies for different inputs. The introduced adaptive feature augmentation, coupled with an adaptive cross-attention mechanism, plays a pivotal role in adaptively exploring and fusing information, yielding superior detection results across a broad spectrum of images. In contrast, previous methods struggle to handle diverse features and patterns, leading to unsatisfactory performance on certain inputs.

In addition to the above quantitative results, Figure 4 presents visual comparisons on several challenging scenes. These testing cases exhibit varied features and patterns, making them difficult for a single unified model to discover salient objects from them effectively. Compared with other methods, our model demonstrates robust performance in handling these challenging cases shown in Figure 4.

3.4 Ablation Study

We conduct a comprehensive series of ablation studies:

1) Training without the proposed adaptive feature augmentation and adaptive cross-attention (**w/o AFA & ACA**); 2) Training solely without the proposed adaptive feature augmentation (**w/o AFA**); 3) Training solely without the adaptive cross-attention mechanism (**w/o ACA**).

It can be inferred from Table 2 and Figure 5 that, omitting both the adaptive feature augmentation (AFA) and the adaptive cross-attention (ACA), denoted as **w/o AFA & ACA**, leads to a significant drop in performance. Besides, abandoning either AFA or ACA also results in performance degradation, underscoring the necessity of incorporating both of the two modules. In summary, incorporating both the proposed adaptive feature augmentation and

Table 2: Ablation studies on the DUTS, DUT-O, and HKU-IS datasets, in terms of Max F-measure (F_{max}), Mean Absolute Error (MAE). The best results are highlighted in **Bold**.

Configuration	DUTS		DUT-O		HKU-IS	
	$F_{max} \uparrow$	MAE $\downarrow$	$F_{max} \uparrow$	MAE $\downarrow$	$F_{max} \uparrow$	MAE $\downarrow$
w/o AFA & ACA	.927	.028	.873	.045	.956	.025
w/o AFA	.934	.024	.878	.042	.963	.020
w/o ACA	.931	.026	.882	.040	.960	.023
Ours	.939	.021	.890	.038	.966	.016

adaptive cross-attention into our dual-branch framework, yields the optimal results, thanks to the ability to adaptively explore and integrate the relevant/valuable information from varying input images with unique patterns/characteristics.

4 Conclusion

In this work, we customized a dual-branch deep learning framework to simplify the original challenging SOD problem into two manageable sub-tasks, *i.e.*, object localization and boundary refinement. A novel adaptive feature augmentation technique was proposed to adaptively identify the potential misleading attributions in the feature space, and enhance the robustness against them. This technique prioritizes preserving features relevant to the detection targets while suppressing the adverse impacts of disruptive or detrimental attributions. A cross-attention module was also introduced to adaptively generate the linear parameters of queries, keys, and values. Extensive experiments were conducted to verify the effectiveness of our design, and its superiority over other state-of-the-art methods. In the future, we aim to extend the potential of our method to other tasks, such as camouflaged object detection and semantic segmentation.

Acknowledgements. This work was supported in part by the Science and Technology Program of Jiangsu Province (BZ2024062), the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (22KLA520007), and Suzhou Planning Project of Science and Technology (2023ss03, SYG2025129, SNG2025010).

Declarations

Funding

Science and Technology Program of Jiangsu Province (BZ2024062); Natural Science Foundation of the Jiangsu Higher Education Institutions of China (22KLA520007); Suzhou Planning Project of Science and Technology (2023ss03, SYG2025129, SNG2025010).

Competing interests

The author(s) declare no competing interests.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Data availability

The data used in this study can be obtained from the following sources:

- 1) The dataset named **DUTS** can be found at: <https://saliencydetection.net/duts/>;
- 2) The dataset named **DUT-O** can be found at: <http://ice.dlut.edu.cn/lu/DUT-OMRON/Homepage.htm>;
- 3) The dataset named **HKU-IS** can be found at: https://github.com/ForawardStar/Datasets/releases/download/HKU-IS_Datasets/HKU-IS.zip;
- 4) The dataset named **ECSSD** can be found at: <https://www.cse.cuhk.edu.hk/leojia/projects/hsaliency/dataset.html>;
- 5) The dataset named **PASCAL-S** is built on the validation set of PASCAL VOC 2010 segmentation challenge, through performing the “free-viewing” task. The link of the validation set of PASCAL VOC 2010 segmentation challenge is: <https://www.robots.ox.ac.uk/~vgg/projects/pascal/VOC/voc2010/index.html>. The PASCAL-S dataset used for salient object detection can be found at: https://github.com/ForawardStar/Datasets/releases/download/SOD_Datasets/PASCAL-S.rar.

Materials availability

Not applicable.

Code availability

Not applicable.

Author contribution

The author contributions are summarized as:

- 1) Yuanbin Fu has made the contribution to the conception of the work, the acquisition, analysis, and drafted the work;
- 2) Jiahao Wang has made the contribution to the conception of the work, and substantively revised the manuscript.

References

- [1] Gurkan, F., Cerkezi, L., Cirakman, O., Günsel, B.: Tdiot: target-driven inference for deep video object tracking. *TIP* **30**, 7938–7951 (2021)
- [2] Mahadevan, V., Vasconcelos, N.: Saliency-based discriminant tracking. In: *CVPR*, pp. 1007–1013 (2009)
- [3] Wu, R., Feng, M., Guan, W., Wang, D., Lu, H., Ding, E.: A mutual learning method for salient object detection with intertwined multi-supervision. In: *CVPR*, pp. 8150–8159 (2019)

- [4] Liu, J., Hou, Q., Cheng, M., Feng, J., Jiang, J.: A simple pooling-based design for real-time salient object detection. In: CVPR, pp. 3917–3926 (2019)
- [5] Chen, S., Tan, X., Wang, B., Hu, X.: Reverse attention for salient object detection. In: ECCV, vol. 11213, pp. 236–252 (2018)
- [6] Wang, W., Lai, Q., Fu, H., Shen, J., Ling, H., Yang, R.: Salient object detection in the deep learning era: an in-depth survey. TPAMI **44**(6), 3239–3259 (2022)
- [7] Pang, Y., Zhao, X., Zhang, L., Lu, H.: Multi-scale interactive network for salient object detection. In: CVPR, pp. 9410–9419 (2020)
- [8] Siris, A., Jiao, J., Tam, G.K.L., Xie, X., Lau, R.W.H.: Scene context-aware salient object detection. In: ICCV, pp. 4136–4146 (2021)
- [9] Piao, Y., Wang, J., Zhang, M., Lu, H.: Mfnet: multi-filter directive network for weakly supervised salient object detection. In: ICCV, pp. 4116–4125 (2021)
- [10] Gu, Y.-C., Gao, S.-H., Cao, X.-S., Du, P., Lu, S.-P., Cheng, M.-M.: Inas: integral nas for device-aware salient object detection. In: ICCV, pp. 4914–4924 (2021)
- [11] Tian, X., Zhang, J., Xiang, M., Dai, Y.: Modeling the distributional uncertainty for salient object detection models. In: CVPR, pp. 19660–19670 (2023)
- [12] Zhang, W., Zheng, L., Wang, H., Wu, X., Li, X.: Saliency hierarchy modeling via generative kernels for salient object detection. In: ECCV, vol. 13688, pp. 570–587 (2022)
- [13] Gao, S., Tan, Y., Cheng, M., Lu, C., Chen, Y., Yan, S.: Highly efficient salient object detection with 100k parameters. In: ECCV, vol. 12351, pp. 702–721 (2020)
- [14] Wei, J., Wang, S., Wu, Z., Su, C., Huang, Q., Tian, Q.: Label decoupling framework for salient object detection. In: CVPR (2020)
- [15] Li, X., Yang, F., Cheng, H., Liu, W., Shen, D.: Contour knowledge transfer for salient object detection. In: ECCV, vol. 11219, pp. 370–385 (2018)
- [16] Zhao, X., Pang, Y., Zhang, L., Lu, H.: Joint learning of salient object detection, depth estimation and contour extraction. TIP **31**, 7350–7362 (2022)
- [17] Wang, Y., Wang, R., Fan, X., Wang, T., He, X.: Pixels, regions, and objects: multiple enhancement for salient object detection. In: CVPR, pp. 10031–10040 (2023)
- [18] Wang, Y., Zhao, X., Hu, X., Li, Y., Huang, K.: Focal boundary guided salient object detection. TIP **28**(6), 2813–2824 (2019)
- [19] Song, X., Guo, F., Zhang, L., Lu, X., Hei, X.: Salient object detection with dual-branch stepwise feature fusion and edge refinement. TCSVT **34**(4), 2832–2844 (2024)

- [20] Zhao, J., Liu, J., Fan, D., Cao, Y., Yang, J., Cheng, M.: Egnets: edge guidance network for salient object detection. In: ICCV, pp. 8778–8787 (2019)
- [21] Zhao, J., Jia, Y., Ma, L., Yu, L.: Recurrent adaptive graph reasoning network with region and boundary interaction for salient object detection in optical remote sensing images. *TGRS* **62**, 1–20 (2024)
- [22] Feng, M., Lu, H., Ding, E.: Attentive feedback network for boundary-aware salient object detection. In: CVPR, pp. 1623–1632 (2019)
- [23] Li, J., Wang, Z., Pan, Z., Liu, Q., Guo, D.: Looking at boundary: siamese densely cooperative fusion for salient object detection. *TNNLS* **34**(7), 3580–3593 (2023)
- [24] Liu, J., Hou, Q., Cheng, M.: Dynamic feature integration for simultaneous detection of salient object, edge, and skeleton. *TIP* **29**, 8652–8667 (2020)
- [25] Yao, Z., Wang, L.: Boundary information progressive guidance network for salient object detection. *TMM* **24**, 4236–4249 (2022)
- [26] Wei, L., Zong, G.: Ega-net: edge feature enhancement and global information attention network for rgb-d salient object detection. *Inf. Sci.* **626**, 223–248 (2023)
- [27] Yang, Z., Soltanian-Zadeh, S., Farsiu, S.: Biconnet: an edge-preserved connectivity-based approach for salient object detection. *PR* **121**, 108231 (2022)
- [28] Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A., Chen, L.: Mobilenetv2: inverted residuals and linear bottlenecks. In: CVPR, pp. 4510–4520 (2018)
- [29] Wu, Y.-H., Liu, Y., Zhan, X., Cheng, M.-M.: P2t: pyramid pooling transformer for scene understanding. *TPAMI* (2022)
- [30] Camuto, A., Willetts, M., Simsekli, U., Roberts, S.J., Holmes, C.C.: Explicit regularisation in gaussian noise injections. In: NeurIPS (2020)
- [31] Lim, S.H., Erichson, N.B., Utrera, F., Xu, W., Mahoney, M.W.: Noisy feature mixup. In: ICLR (2022)
- [32] Cao, Z., Long, M., Wang, J., Yu, P.S.: Hashnet: deep learning to hash by continuation. In: ICCV, pp. 5609–5618 (2017)
- [33] Zhang, M., Liu, S., Zhu, Y., Bai, Z., Lin, J.: Seg-hashnet: semantic segmentation based unsupervised hashing. In: ICIP, pp. 2371–2375 (2020)
- [34] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: transformers for image recognition at scale. In: ICLR (2021)

- [35] Chiu, Y., Chen, G., Chen, Z., Peng, Y.: Abc-former: auxiliary bimodal cross-domain transformer with interactive channel attention for white balance. In: CVPR, pp. 21258–21266 (2025)
- [36] Yan, X., Cai, Y., Wang, Q., Zhou, Y., Huang, W., Yang, H.: Long video diffusion generation with segmented cross-attention and content-rich video data curation. In: CVPR, pp. 3184–3194 (2025)
- [37] Lai, H., Yao, Q., Jiang, Z., Wang, R., He, Z., Tao, X., Zhou, S.K.: Carzero: cross-attention alignment for radiology zero-shot classification. In: CVPR, pp. 11137–11146 (2024)
- [38] Liu, B., Wang, C., Cao, T., Jia, K., Huang, J.: Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In: CVPR, pp. 7817–7826 (2024)
- [39] Milletari, F., Navab, N., Ahmadi, S.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: International Conference on 3D Vision, pp. 565–571 (2016)
- [40] Xie, S., Tu, Z.: Holistically-nested edge detection. IJCV **125**(1-3), 3–18 (2017)
- [41] Zhao, X., Pang, Y., Zhang, L., Lu, H., Zhang, L.: Suppress and balance: a simple gated network for salient object detection. In: ECCV, vol. 12347, pp. 35–51 (2020)
- [42] Wei, J., Wang, S., Wu, Z., Su, C., Huang, Q., Tian, Q.: Label decoupling framework for salient object detection. In: CVPR, pp. 13022–13031 (2020)
- [43] Liu, N., Zhang, N., Wan, K., Shao, L., Han, J.: Visual saliency transformer. In: ICCV, pp. 4702–4712 (2021)
- [44] Yun, Y.K., Lin, W.: Selfreformer: self-refined network with transformer for salient object detection (2022)
- [45] Wu, Y., Liu, Y., Zhang, L., Cheng, M., Ren, B.: Edn: salient object detection via extremely-downsampled network. TIP **31**, 3125–3136 (2022)
- [46] Ma, M., Xia, C., Xie, C., Chen, X., Li, J.: Boosting broader receptive fields for salient object detection. TIP **32**, 1026–1038 (2023)
- [47] Liu, W., Shen, X., Pun, C., Cun, X.: Explicit visual prompting for low-level structure segmentations. In: CVPR, pp. 19434–19445 (2023)
- [48] Yun, Y.K., Lin, W.: Towards a complete and detail-preserved salient object detection. TMM **26**, 4667–4680 (2024)
- [49] Gao, S., Zhang, P., Yan, T., Lu, H.: Multi-scale and detail-enhanced segment anything model for salient object detection. In: ACM MM, pp. 9894–9903 (2024)

- [50] Zhu, J., Qin, X., Elsaddik, A.: Dc-net: divide-and-conquer for salient object detection. *PR* **157**, 110903 (2025)
- [51] Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: *ICLR* (2015)
- [52] Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: *CVPR*, pp. 248–255 (2009)
- [53] Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: *CVPR*, pp. 3796–3805 (2017)
- [54] Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.: Saliency detection via graph-based manifold ranking. In: *CVPR*, pp. 3166–3173 (2013)
- [55] Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: *CVPR*, pp. 5455–5463 (2015)
- [56] Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: *CVPR*, pp. 1155–1162 (2013)
- [57] Li, Y., Hou, X., Koch, C., Rehg, J.M., Yuille, A.L.: The secrets of salient object segmentation. In: *CVPR*, pp. 280–287 (2014)
- [58] Margolin, R., Zelnik-Manor, L., Tal, A.: How to evaluate foreground maps. In: *CVPR*, pp. 248–255 (2014)