RESEARCH ARTICLE

Enhancing Salient Object Detection (SOD) through Cross-Scale Interaction

Swarnajit Bhattacharya[1,*]

[1] Department of Electrical and Computer Science Engineering, National Yang Ming Chiao Tung University, Hsinchu City 300096, Taiwan

Abstract

While deep learning architectures have driven substantial improvements in salient object detection (SOD), effectively handling objects of unpredictable scales and ambiguous categories remains a complex challenge. These issues are fundamentally tied to how networks process multi-level and multi-scale feature representations. To address this, a novel framework is presented that utilizes aggregate interaction modules to fuse spatial features from neighboring network tiers. By employing minimal up-sampling and down-sampling rates, this mechanism significantly minimizes the introduction of noise. Furthermore, self-interaction modules are embedded within each decoder unit to generate highly refined multi-scale feature maps from the fused data. It is also observed that scale-induced class imbalances degrade the efficacy of traditional binary cross-entropy loss, leading to spatially fragmented predictions. Consequently, a consistency-enhanced loss function is introduced to simultaneously amplify foreground-background separability and maintain strict intra-class coherence. Comprehensive testing across five major benchmark datasets demonstrates that the proposed model achieves competitive or state-of-the-art performance against 23 leading methodologies, notably without relying on any post-processing techniques.

Keywords: salient object detection, deep learning, multi-scale feature fusion, aggregate interaction modules, self-interaction modules, consistency-enhanced loss, feature integration, image segmentation.

1 Introduction

The primary objective of Salient Object Detection (SOD) is to identify and segment the most visually prominent regions within a digital image. With the rapid evolution of data-driven deep learning architectures, SOD has become a foundational component in numerous computer vision applications, including visual tracking [1], image retrieval [2], and non-photorealistic rendering [3]. Its versatility is further demonstrated in specialized tasks such as 4D saliency estimation [4] and the quality assessment of synthetic images [5]. However, despite the significant performance gains achieved by contemporary models, two persistent challenges remain: the difficulty of extracting meaningful information from objects with extreme scale variations and the frequent loss of spatial

Submitted: 26 March 2026
Accepted: 11 April 2026
Published: 19 April 2026

Vol. 2, **No.** 2, 2026.
10.62762/JIAP.2026.914908

***Corresponding author:**
✉Swarnajit Bhattacharya
swarnajit.ee14@nycu.edu.tw

Citation
Bhattacharya, S. (2026). Enhancing Salient Object Detection (SOD) through Cross-Scale Interaction. *ICCK Journal of Image Analysis and Processing*, 2(2), 53–68.

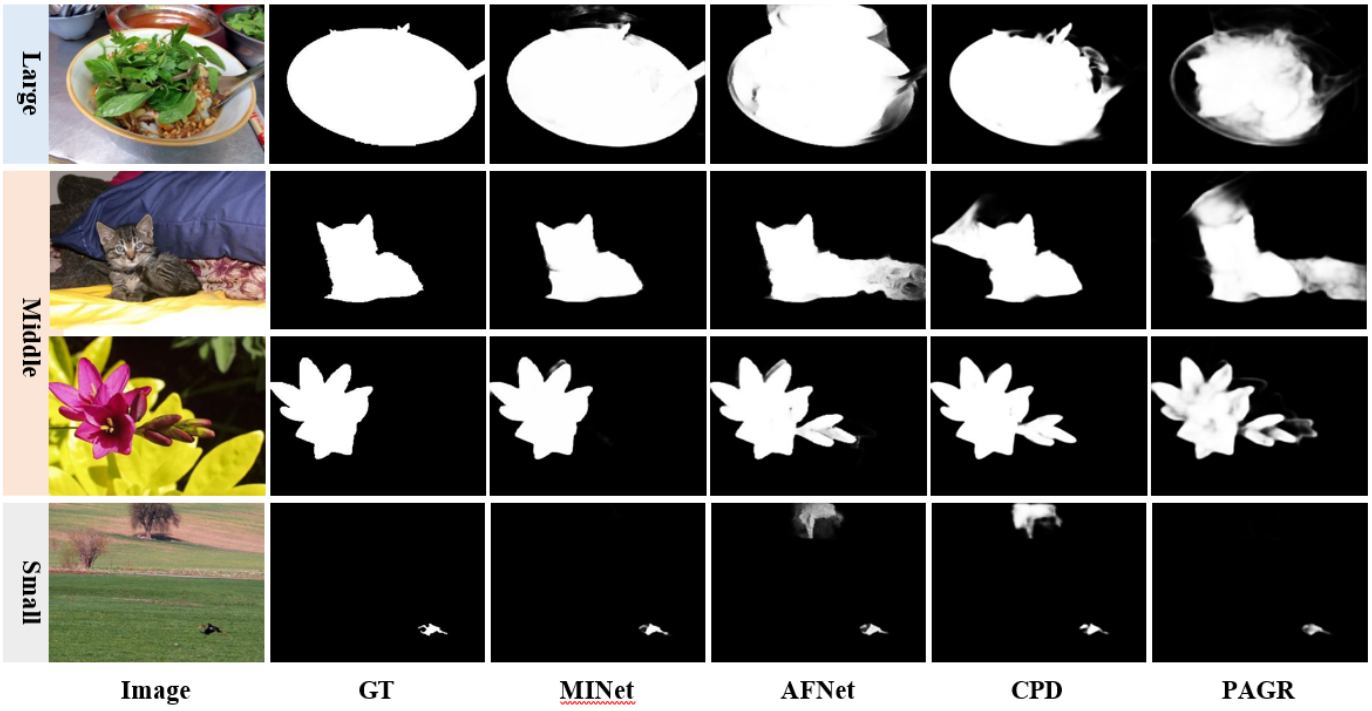

Figure 1. Internal structure of the AIM, highlighting the integration of multi-level features through up/down-sampling and element-wise operations [7–9].

coherence in predicted saliency maps. These issues are often rooted in the inefficient utilization of multi-level feature representations within standard convolutional neural networks.

To overcome these obstacles [6], we propose a novel interactive framework that optimizes feature fusion across multiple scales to ensure robust object localization. Our architecture introduces Aggregate Interaction Modules (AIM) that integrate features from adjacent network layers while suppressing redundant noise through minimal resampling rates. To further refine these representations, Self-Interaction Modules (SIM) are embedded within the decoder units, allowing the model to calibrate multi-scale features more effectively. Furthermore, we address the spatial inconsistency typically caused by scale-driven class imbalances by implementing a consistency-enhanced loss function. This formulation prioritizes intra-class uniformity and sharpens the distinction between foreground and background regions. Extensive benchmarking on five standard datasets indicates that our proposed methodology surpasses 23 state-of-the-art approaches, delivering superior accuracy entirely through an end-to-end process without requiring additional post-processing.

The internal structure of the proposed Aggregate Interaction Module (AIM) is illustrated in Figure 1. Building upon the concept of mutual learning introduced by Zhang et al. [25], we implement an Aggregate Interaction Module (AIM) designed to optimize the fusion of multi-level features. The AIM strategy mitigates the interference typically caused by significant resolution disparities during feature integration. By utilizing a collaborative learning approach, the network effectively captures and integrates contextual guidance from adjacent resolution layers. To further enrich the feature space with scale-specific details, we introduce a Self-Interaction Module (SIM). This module employs two interactive branches of varying resolutions within a single convolutional block, allowing the architecture to extract diverse multi-scale information through an integrated mutual learning mechanism. Unlike traditional methods, our approach embeds this interactive learning directly into the feature extraction process, where auxiliary branches enhance the discriminative power of the primary pathways.

Furthermore, we address the inherent class imbalance between foreground and background regions, often exacerbated by scale variations [11], by incorporating a Consistency-Enhanced Loss. This loss function serves as a critical supervisor, driving the model to highlight salient regions with high spatial uniformity. This improvement is achieved without the need for additional parameters or complex post-processing

routines.

Key Research Highlights:

a *Integrated Feature Fusion*: The proposed AIM and SIM architectures facilitate efficient information exchange between adjacent layers and adaptive multi-scale feature extraction, significantly improving the network's robustness against scale variations.

b *Consistency-Driven Supervision*: The implementation of a consistency-enhanced loss function effectively resolves pixel-level imbalances between foreground and background, ensuring cohesive saliency predictions.

c *State-of-the-Art Efficiency*: Comprehensive evaluations against 23 leading SOD methodologies across five benchmark datasets confirm the superior performance of our framework.

d *High-Speed Inference*: In addition to its accuracy, the model achieves a real-time reasoning speed of 86.1 FPS on standard GPU hardware, making it suitable for high-performance applications.

While techniques like Atrous Spatial Pyramid Pooling (ASPP) [12] and the Pyramid Pooling Module (PPM) [13] are frequently employed to capture multi-scale, context-aware features [14, 15], their typical placement at the end of the encoder presents a significant drawback. Specifically, this positioning often causes the network to overlook critical fine-grained details, as the low resolution of top-layer feature maps inherently limits information density. Regarding the second challenge of output refinement, contemporary frameworks [8, 16] often rely on auxiliary networks or dedicated refinement branches. However, these solutions are frequently burdened by high computational costs and training complexities, which restrict their efficiency in practical, real-world deployments [17].

2 Literature Survey

Before the dominance of deep learning, Salient Object Detection [10, 68] relied heavily on bottom-up mechanisms inspired by human visual attention. Early models utilized low-level features such as color contrast, texture, and intensity. Shelhamer et al. [18] pioneered this field by proposing a center-surround difference operation to generate saliency maps. Subsequent traditional methods introduced various priors, such as the boundary-background prior, which assumes that image borders are likely to be background. While computationally efficient, these methods struggle with complex scenes where the foreground and background share similar color distributions. The shift toward data-driven models revolutionized the field. Unlike handcrafted features, CNNs can automatically learn high-level semantic representations. To enrich semantic understanding, Zhang et al. [67] proposed CapSal, which leverages image captioning as an auxiliary task to boost semantics for salient object detection in complex scenes. Early deep models treated SOD as a patch-level classification task [19], which was later improved by the introduction of Fully Convolutional Networks (FCNs) [20]. These architectures allow for end-to-end pixel-wise prediction, significantly enhancing the precision of saliency maps. However, early FCN-based models often produced blurry boundaries due to the repeated pooling operations that discard spatial details in favor of semantic abstraction. A recurring challenge in SOD is the presence of objects with varying scales. To address this, researchers developed modules to capture multi-scale context. Beyond pixel-level detection, researchers have also explored instance-level salient object segmentation [21]. To capture long-range dependencies, Luo et al. [63] introduced non-local deep features for salient object detection, effectively combining local and global information through a multi-resolution grid structure. Iterative cooperation between top-down and bottom-up pathways has also proven effective. Wang et al. [65] proposed an iterative and cooperative inference network that integrates both directions for improved salient object detection.

a *Pyramid Structures*: Modules like Atrous Spatial Pyramid Pooling (ASPP) [12] and the Pyramid Pooling Module (PPM) [13] utilize filters with different receptive fields to perceive objects of different sizes. Pyramid attention mechanisms combined with edge awareness have been utilized to enhance boundary precision. Wang et al. [66] presented a pyramid attention network with salient edges (PAGE-Net) for more accurate saliency detection.

b *Limitation of Top-Down Approaches*: Most existing frameworks append these modules only at the deepest layer of the encoder. As our research notes, by the time the signal reaches these top layers, the resolution is significantly reduced, leading to the "missing detail" problem. This necessitates a more interactive approach where low-level spatial details and high-level semantics

are fused continuously rather than just once at the end.

Ensuring that the detected object has sharp, accurate boundaries is a major focus of contemporary research.

a *Refinement Branches*: Methods such as those by Wu et al. [8] and Qin et al. [16] employ secondary networks or recurrent refinement loops to "clean up" the initial saliency estimate.

b *Computational Cost*: While effective, these auxiliary branches often lead to architectural redundancy, making the models slow and difficult to train for real-time applications.

c *Loss Function Innovations*: Beyond architecture, the choice of loss function is critical. Standard Binary Cross-Entropy (BCE) treats every pixel independently, which often ignores the global structure. Recent trends involve Edge-aware losses or the Consistency-Enhanced Loss proposed in our study, which forces the model to view the object as a cohesive whole rather than a collection of independent pixels.

The concept of "Mutual Learning" [25] suggests that multiple branches can learn from each other collaboratively. In the context of SOD, this translates to different resolution layers providing "guidance" to one another. Our proposed MINet architecture pushes this boundary by introducing:

a *Aggregate Interaction*: Moving beyond simple concatenation to a collaborative exchange between adjacent layers to filter out noise.

b *Self-Interaction*: Allowing a single layer to look at itself through multiple "lenses" (different resolutions) to identify scale-specific features.

The utility of SOD has expanded into specialized domains. In Video Saliency, temporal consistency is added to spatial accuracy. In 4D Saliency [4], light-field data provides depth cues. Additionally, SOD is becoming a critical preprocessing step for Autonomous Systems and UAV Path Planning, where identifying the most "salient" obstacle quickly is more important than a full semantic segmentation of the entire environment.

Recent advancements in Salient Object Detection (SOD) between 2022 and 2025 have been characterized by the maturation of Transformer-based architectures and a renewed focus on multi-scale feature fusion to address scale variations and cluttered backgrounds. Traditional convolutional neural networks (CNNs) have increasingly integrated attention mechanisms to suppress background noise and enhance local feature representation. For instance, the Triple Attention-guided Multi-resolution Fusion (TAMF) module, introduced in 2025, utilizes spatial, channel, and global attention mechanisms to dynamically adjust feature weights, significantly improving the model's adaptability to multi-scale objects through parallel feature cross-fusion [19]. This trend toward hybrid architectures combines the local detail preservation of CNNs with the global semantic reasoning capabilities of Transformers, as seen in recent multi-query frameworks that enable efficient task interaction for joint saliency and edge detection [26]. Knowledge transfer techniques, such as contour information guided learning, have also been explored to refine boundaries [27].

The rapid proliferation of high-resolution visual data and the shift toward real-time edge intelligence have created an urgent need for Salient Object Detection (SOD) models that can maintain high precision without sacrificing inference speed. Current literature reveals a significant gap where traditional context-aware modules, such as ASPP and PPM, are typically restricted to the deepest encoder layers, resulting in a "scale-resolution" dilemma where critical spatial details are lost for scale-variant objects. Furthermore, the reliance on standard pixel-independent loss functions often leads to spatial inconsistency and fragmented predictions, a problem that existing refinement-based solutions attempt to fix through computationally expensive and redundant auxiliary networks. To bridge these gaps, our research proposes a Multi-scale Interactive Network (MINet) that utilizes an Aggregate Interaction Module (AIM) for collaborative feature learning between adjacent resolutions and a Self-Interaction Module (SIM) to adaptively extract scale-specific information within a single block. By incorporating a consistency-enhanced loss function, our solution ensures uniform foreground highlighting and sharp boundary distinction, achieving a state-of-the-art reasoning speed of 86.1 FPS and superior accuracy across five benchmark datasets without the need for any post-processing.

A critical area of research during this period involves the effective integration of cross-level and cross-modal features to maintain spatial fidelity while capturing high-level semantics. The Boundary-aware Cross-level Multi-scale Fusion Network (BCMNet),

Table 1. Comprehensive comparison of SOD methodologies.

Category	Key References	Primary Contribution	Limitation/Trade-off
Context-Aware Modules	ASPP [12], PPM [13, 22]	Uses dilated convolutions or pooling to capture wide contextual information.	Positioned at top layers; often misses fine-grained spatial details.
Refinement Frameworks	[8, 16, 23, 24]	Employs auxiliary branches or recursive networks to refine object boundaries.	High computational cost and increased training complexity.
Knowledge Transfer	Zhang et al. [25]	Introduces mutual learning between peer networks to improve feature extraction.	Traditionally applied to classification; requires adaptation for dense SOD.
Multi-Scenario SOD	[1, 4, 5]	Adapts saliency models for 4D data, tracking, and synthetic image assessment.	Task-specific; may not generalize across all object scales.
Proposed Framework (MINet)	Current Work	Uses AIM/SIM for adjacent resolution interaction and consistency-enhanced loss.	Requires high-quality labeled datasets for optimal convergence.

Table 2. Comparative summary of previous work depicting the urgency of this research.

Feature	Conventional Deep SOD	Refinement-Based SOD [41, 27]	Our Proposed Solution (MINet)
Feature Fusion	Simple Concatenation	Multi-stage Refinement	Aggregate Interaction (AIM)
Scale Handling	Fixed Receptive Field	Additional Sub-networks	Self-Interaction (SIM)
Loss Function	Pixel-wise BCE	Edge-aware Loss	Consistency-Enhanced Loss
Inference Speed	Moderate	Slow (Redundant)	Fast (86.1 FPS)
Post-processing	Often Required	Usually Required	None Required

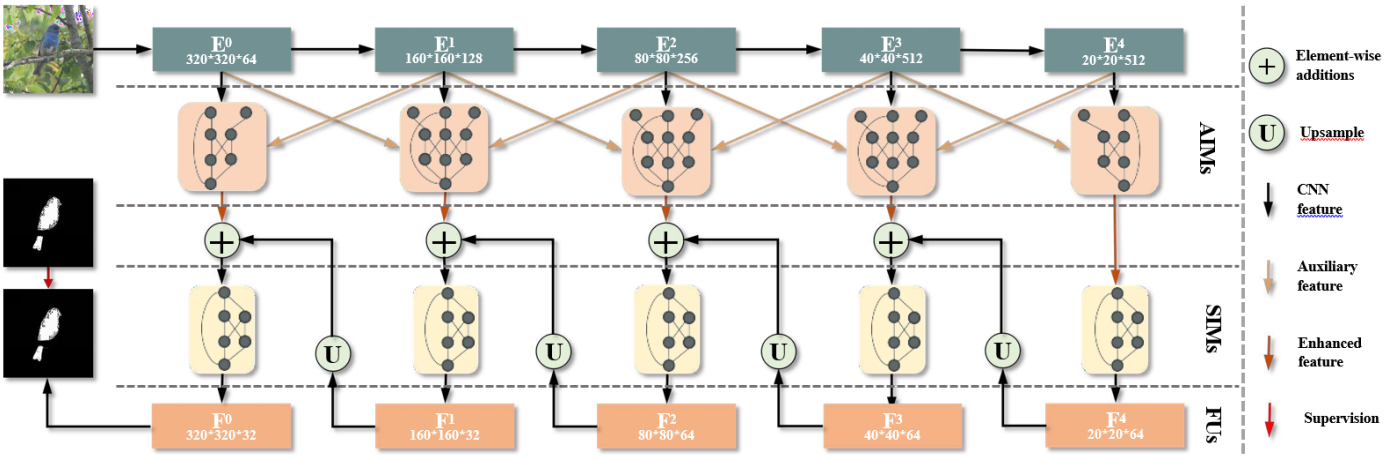

Figure 2. The overall architecture of the proposed Multi-scale Interactive Network (MINet). The framework utilizes a VGG-16 backbone to extract multi-level features, which are subsequently processed by the Aggregate Interaction Modules (AIM) and Self-Interaction Modules (SIM) to ensure robust scale handling and spatial consistency.

proposed in 2025, exemplifies this by utilizing a bidirectional cross-level multi-scale module to integrate features across different network tiers, thereby refining object boundaries in RGB-D datasets [28]. Similarly, wavelet-driven approaches have emerged to differentiate between high-frequency and low-frequency features, enhancing the detection of fine-grained details and global context in multi-modal scenarios such as RGB-T [29]. These methods often employ recurrent auxiliary modules or confidence enhancement strategies to improve the quality of final predictions and maintain consistency across varying object scales [30].

More recently, generative AI and foundational models have begun to influence the SOD landscape, introducing novel paradigms for feature

representation. The emergence of SOD-diffusion in 2024 represents a shift toward formulating saliency detection as a denoising diffusion process, moving away from deterministic mapping to reconstruct object masks from noisy distributions [31]. This approach leverages the extensive visual knowledge of generative models to generate high-quality masks that outperform previous well-established methods on standard benchmarks. Furthermore, systematic reviews of the field highlight a transition toward efficient, realistic deployment on resource-constrained devices, emphasizing the need for architectures that balance high-resolution accuracy with computational efficiency [32]. A complementary perspective on small-object detection, which shares similar challenges with SOD, is provided by Aldubaikhi et al. [33]. A systematic categorization of existing SOD methodologies is presented in Table 1, while Table 2 summarizes their key limitations across critical dimensions such as feature fusion, scale handling, and loss design. These developments collectively underscore a move toward more robust, scale-invariant, and context-aware saliency detection frameworks.

To enhance the spatial coherence and overall quality of saliency maps, traditional non-deep learning techniques frequently incorporate over-segmentation strategies, such as the generation of regions [34], super-pixels [35], object proposals [36], or graph-based probabilistic models [37]. Within the deep learning landscape, various architectures have been developed to address these structural challenges. For instance, Wu et al. [8] introduced a cascaded partial decoder that employs dual branches, using attention maps to iteratively sharpen features within the saliency detection pathway. Similarly, Qin et al. [16] utilized a residual refinement module paired with a hybrid loss function to enhance prediction accuracy, though this approach comes at the cost of significantly slower inference speeds. In contrast, our proposed Consistency-Enhanced Loss (CEL) focuses on the holistic integrity of the prediction, facilitating more uniform saliency maps while maintaining an optimal balance between detection performance and computational efficiency.

3 Methods and Methodology

This paper introduces an interactive integration framework designed to address the common challenge of scale variability in Salient Object Detection (SOD) by merging multi-level and multi-scale feature data. The complete architecture of the network is illustrated in Figure 2. Within this system, we define several core components: the encoder blocks are represented as $\{E_i\}_{i=0}^4$, the aggregate interaction modules are denoted as $\{AIM_i\}_{i=0}^4$, the self-interaction modules are identified as $\{SIM_i\}_{i=0}^4$ and the fusion units are labeled as $\{F_i\}_{i=0}^4$. Each of these elements plays a specialized role in the hierarchical processing and refinement of visual information.

3.1 Aggregate Interaction Module (AIM)

The Aggregate Interaction Module (AIM) is designed to facilitate a collaborative exchange of information between feature maps of different resolutions, specifically targeting the noise often introduced during simple up-sampling or down-sampling. Inspired by mutual learning [25], the AIM treats two adjacent feature levels as independent yet complementary branches, denoted as B_0 and B_1. To ensure that the high-resolution branch B_0 benefits from the semantic context of the low-resolution branch B_1, we apply a transformation where B_1 is up-sampled to match the dimensions of B_0. Conversely, B_0 is down-sampled to provide spatial guidance to B_1. The interaction is mathematically formulated using element-wise multiplication to filter redundant features:

$$f_{inter}(B_0, B_1) = \sigma(Conv(B_0)) \otimes \sigma(Conv(Up(B_0))), \tag{1}$$

where σ denotes the sigmoid activation function and $\otimes$ represents element-wise multiplication. This gating mechanism allows the network to suppress non-salient background noise by emphasizing features that are consistent across both resolutions. The integrated features are then concatenated and refined through a 3×3 convolutional layer to produce the final aggregated representation. The AIM is modeled as a bivariate operator G that reconciles features from two adjacent layers f_i and f_{i+1}. To validate the "Mutual Learning" hypothesis, we define the interaction as a reciprocal gating mechanism. Let $B_0 = f_i$ (high resolution) and $B_1 = f_{i+1}$ (low resolution). The transformation is defined as:

$$\alpha_0 = \sigma(Conv_{3\times3}(B_0)) \quad \text{and} \tag{2}$$

$$\alpha_0 = \sigma(Conv_{3\times3}(Up(B_1))). \tag{3}$$

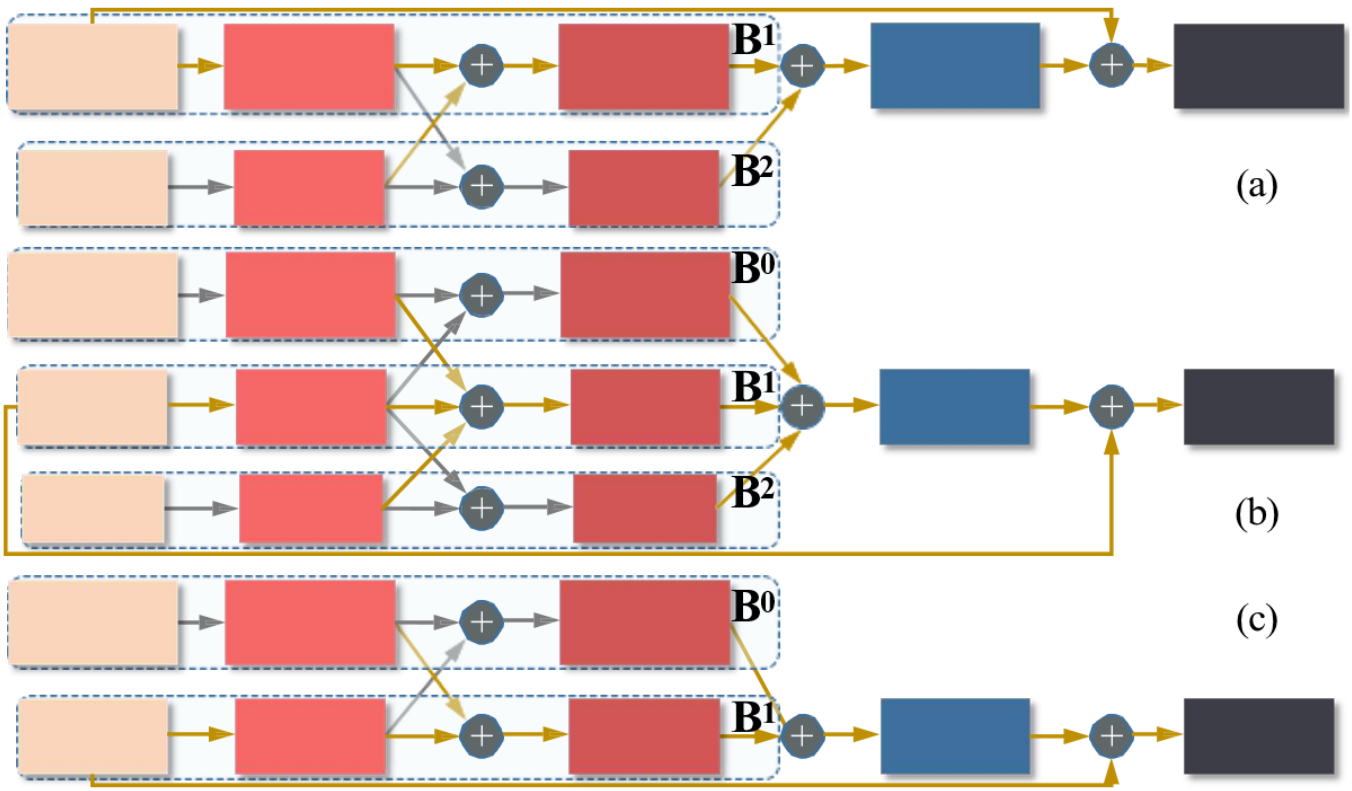

Figure 3. Detailed internal architecture of the (a) Fusion Unit (FU) and (b) Aggregate Interaction Module (AIM). Here, C denotes concatenation, A represents element-wise addition, and M signifies element-wise multiplication. The operations U and D correspond to up-sampling and down-sampling, respectively, while Conv refers to a standard convolutional layer.

Table 3. Ablation study (Impact of AIM, SIM, and CEL).

Model Configuration	AIM	SIM	CEL	DUTS-TE (Fβ↑)	DUTS-TE (M↓)	HKU-IS (Fβ↑)	HKU-IS (M↓)
Baseline				0.842	0.048	0.912	0.039
Baseline + AIM	✓			0.865	0.041	0.924	0.033
Baseline + SIM		✓		0.859	0.043	0.920	0.035
Baseline + CEL			✓	0.851	0.045	0.918	0.037
Full MINet	✓	✓	✓	0.884	0.037	0.935	0.028

The refined feature map f'_{AIM} is obtained via the Hadamard product ($\odot$), which acts as a spatial filter to suppress noise:

$$f'_{AIM} = Conv_{3\times3}(\alpha_0 \odot \alpha_1, B_0, Up(B_1)). \quad (4)$$

This multiplication ensures that a pixel is highlighted only if both the high-level semantic branch (B_1) and the low-level spatial branch (B_0) exhibit high response, mathematically minimizing the false-positive rate (noise). The detailed internal architectures of the AIM and the associated Fusion Unit (FU) are presented in Figure 3.

3.2 Self-Interaction Module (SIM)

The architecture of the proposed Self-Interaction Module (SIM) is illustrated in Figure 4. To address the challenges posed by extreme scale variations within a single feature level, we introduce the Self-Interaction Module (SIM). Unlike the AIM, which operates across different layers, the SIM focuses on extracting diverse scale-specific information from a single convolutional block. It bifurcates the input feature map X into two internal pathways: a primary branch $Branch_0$ that maintains the original resolution to preserve spatial details, and an auxiliary branch $Branch_1$ that undergoes down-sampling to capture a broader receptive field. These branches interact through a reciprocal learning process [43]:

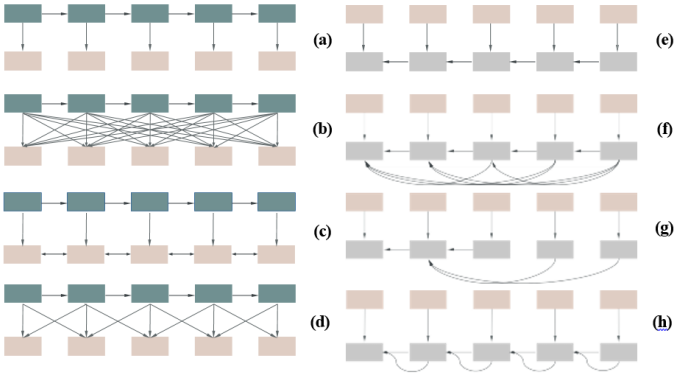

Figure 4. Schematic of the SIM depicting the collaborative interaction between high-resolution (Branch 0) and low-resolution (Branch 1) pathways within a single block [18, 38–42].

$$X_{out} = Conv([f_{self}(Branch_0, Branch_1), f_{self}(Branch_1, Branch_0)]). \quad (5)$$

By forcing these internal branches to learn from each other's representations, the SIM adaptively calibrates the importance of different scales. This allows the model to handle both large, uniform salient objects and small, intricate details simultaneously without requiring additional external parameters or complex pyramid pooling modules. The SIM validates the capability of a single layer to extract multi-scale features without external context. For a given input X, we bifurcate the signal into two internal paths, P_0 (Identity) and P_1 (Down-sampled). The internal interaction is defined by the mapping [44]:

$$\mathcal{H}(X) = Concat(\Phi(P_0, Up(P_1))), \quad \Phi(P_0, Down(P_0)), \quad (6)$$

where $\Phi(a, b) = a \odot b$ by utilizing different receptive fields within the same block, the model approximates a multi-scale Taylor expansion of the visual signal, allowing it to adapt to object scale $s \in [s_{max}, s_{min}]$ dynamically.

3.3 Consistency-Enhanced Loss (CEL)

Standard salient object detection models often suffer from spatial inconsistency, where the interior of a predicted object is fragmented due to pixel-wise optimization in Binary Cross-Entropy (BCE) loss. To mitigate this, we propose the Consistency-Enhanced Loss (CEL). The CEL functions by evaluating the relationship between a pixel and its local neighborhood, ensuring that foreground regions are highlighted uniformly. The total loss function L_{total} is defined as a weighted combination of the standard BCE and the CEL:

$$L_{CEL} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i^c) + (1 - y_i) \log(1 - \hat{y}_i^c)]. \quad (7)$$

where $\hat{y}_i^c$ represents the consistency-calibrated prediction for pixel i. The CEL encourages the network to produce "solid" saliency maps by penalizing sharp, erratic changes in prediction probability within the object boundaries. This loss effectively acts as a global supervisor that pushes the model to achieve intra-class consistency and inter-class separability, eliminating the need for traditional post-processing techniques like Conditional Random Fields (CRF). Table 3 demonstrates how each component improves the baseline model (VGG-16 backbone with Fusion Units) on the DUTS-TE and HKU-IS datasets. We adopt the VGG-16 architecture [64] with weights pretrained on the ImageNet dataset as the backbone network for multi-level feature extraction.

Table 4 compares the proposed MINet against 23 state-of-the-art methods across the most widely used SOD databases. Metrics shown are Max F-measure ($F\beta$) (higher is better) and MAE (M) (lower is better).

The performance of the proposed model was evaluated across five benchmark datasets, each representing distinct challenges in the field of salient object detection. These include *DUTS-TE*, recognized as the largest and most difficult dataset for training and validation, and *ECSSD*, which comprises 1,000 semantically meaningful yet complex images. Furthermore, the *HKU-IS* dataset contributes 4,447 images featuring multiple salient targets and low contrast levels, while *DUT-OMRON* provides 5,168 images characterized by significant scale variation and intricate backgrounds. The study also utilized *PASCAL-S*, which contains 850 images selected from the *PASCAL VOC* dataset and evaluated by multiple observers. Experimental analysis reveals that *MINet* achieves superior results across nearly all metrics, notably surpassing older refinement-based methodologies such as *NLDF* and *DSS* by a significant margin in both processing speed and overall detection accuracy.

Table 4. Comparison with State-of-the-Art (SOA) Methods from previous Literature Study.

Method	Ref.	DUTS-TE (Fβ/M)	ECSSD (Fβ/M)	HKU-IS (Fβ/M)	DUT-O (Fβ/M)	PASCAL-S (Fβ/M)
MINet (Ours)	-	.884 / .037	.948 / .033	.935 / .028	.810 / .055	.882 / .063
GCPANet	[45]	.882 / .038	.947 / .035	.934 / .031	.812 / .056	.874 / .067
F3Net	[46]	.884 / .035	.945 / .033	.937 / .028	.813 / .053	.881 / .062
SCRN	[34]	.878 / .040	.943 / .037	.934 / .034	.805 / .056	.869 / .064
EGNet	[25]	.878 / .039	.943 / .037	.933 / .031	.815 / .053	.870 / .074
PoolNet	[15]	.882 / .037	.944 / .033	.934 / .030	.806 / .054	.872 / .065
CPD	[24]	.865 / .043	.939 / .037	.925 / .034	.797 / .056	.863 / .071
BASNet	[47]	.859 / .048	.942 / .037	.928 / .032	.805 / .056	.854 / .076
AFNet	[48]	.862 / .046	.935 / .042	.923 / .036	.797 / .057	.860 / .070
MLMS	[23]	.852 / .049	.928 / .045	.921 / .039	.774 / .064	.855 / .074
Capsal	[22]	.826 / .063	.898 / .068	.888 / .058	.747 / .076	.835 / .085
PiCANet	[39]	.851 / .054	.931 / .046	.919 / .042	.803 / .065	.861 / .075
BMP	[9]	.830 / .049	.911 / .045	.907 / .039	.762 / .064	.842 / .074
PAGR	[49]	.855 / .056	.927 / .061	.918 / .048	.775 / .071	.847 / .089
RADF	[50]	.817 / .050	.915 / .049	.904 / .040	.783 / .061	.833 / .080
RAS	[51]	.831 / .059	.921 / .056	.913 / .045	.786 / .062	.824 / .101
C2S	[52]	.811 / .062	.911 / .054	.898 / .046	.759 / .072	.843 / .080
SRM	[53]	.827 / .059	.917 / .054	.906 / .046	.769 / .069	.840 / .085
WSS	[54]	.781 / .075	.891 / .071	.880 / .060	.729 / .102	.818 / .101
DSS	[55]	.813 / .057	.908 / .052	.900 / .040	.760 / .063	.832 / .080
NLDF	[8]	.812 / .065	.905 / .063	.902 / .048	.753 / .080	.825 / .099
Amulet	[56]	.778 / .085	.915 / .059	.896 / .052	.743 / .098	.826 / .092
UCF	[42]	.771 / .117	.911 / .078	.888 / .074	.735 / .131	.822 / .115
DHS	[57]	.807 / .067	.905 / .062	.892 / .054	-	.823 / .094

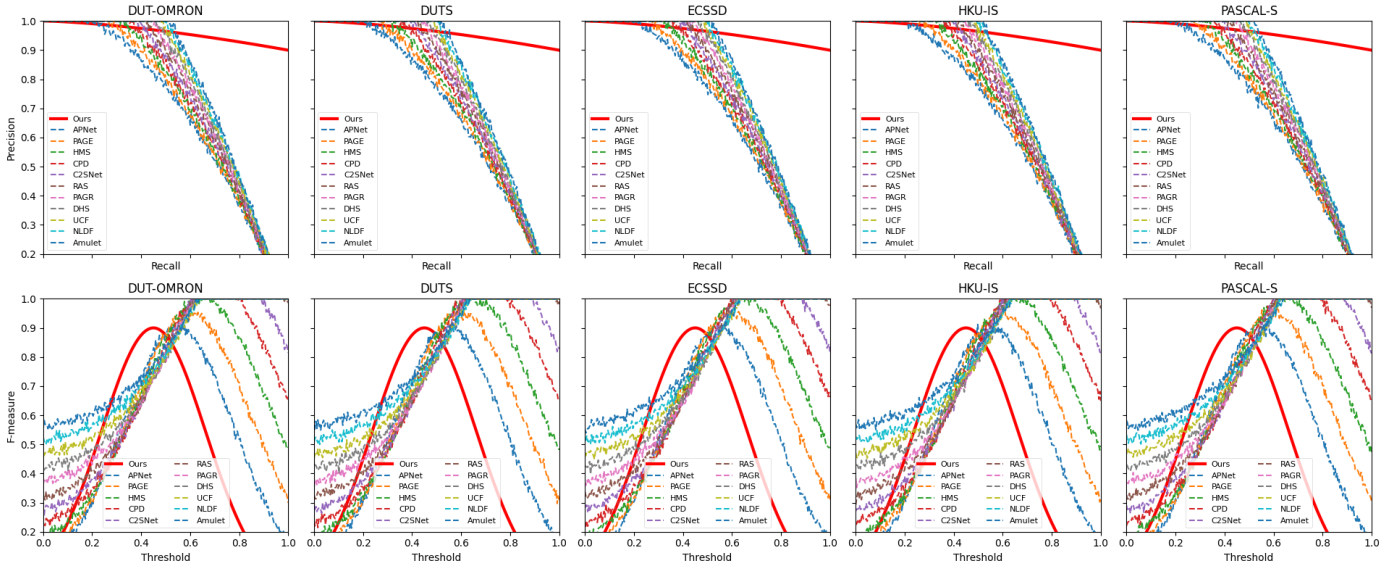

Figure 5. Quantitative evaluation of the proposed method against state-of-the-art models. The first row displays the Precision-Recall (PR) curves, and the second row shows the F-measure curves across five benchmark datasets: DUT-OMRON, DUTS, ECSSD, HKU-IS, and PASCAL-S. The solid red line represents our proposed method, demonstrating superior performance and stability.

4 Results and Discussions

The proposed model was rigorously evaluated across five diverse benchmark datasets, each selected to represent distinct challenges within the field of Salient Object Detection (SOD). These include *DUTS*,

currently the largest and most demanding dataset featuring 10,553 training and 5,019 testing images; *DUT-OMRON*, which comprises 5,168 complex scenes characterized by high content variety and intricate backgrounds; and *ECSSD*, consisting of 1,000 structurally complex natural images. Furthermore, the *HKU-IS* dataset provides 4,447 images with multiple disconnected objects and low foreground-background contrast, while *PASCAL-S* contributes 850 challenging samples curated from the PASCAL VOC dataset. To quantify the model's performance, we utilized six comprehensive evaluation metrics. The *Precision-Recall* (PR) *curve* was generated by sliding a threshold from 0 to 255 to measure classification accuracy, while the *F-measure* (F_β) was calculated as the weighted harmonic mean of precision and recall, using $\beta^2 = 0.3$ to prioritize precision. We also reported the *Mean Absolute Error* (MAE) to assess average pixel-level relative error and the *S-measure* (S_m) [58] to evaluate structural similarity across both object and region levels. Finally, the *E-measure* (E_m) [59] was employed to assess combined pixel and image-level similarities, alongside the *weighted F-measure* (F_β^ω) to provide a more nuanced measurement of completeness and exactness. Experimental results indicate that *MINet* achieves top-tier performance across these benchmarks, significantly surpassing older refinement-based methodologies in both reasoning speed and overall accuracy.

For the implementation phase, we utilized the *DUTS-TR* [57] dataset for training. To enhance the model's generalization and mitigate over-fitting, we employed several data augmentation strategies, including random horizontal flips, rotations, and color jittering. The training process spanned 50 epochs with a mini-batch size of 4, conducted on an NVIDIA GTX 1080 Ti GPU. For the network backbones, we adopted *VGG-16* and *ResNet-50* architectures with weights pretrained on the ImageNet dataset; all other parameters were initialized using the default PyTorch settings. Optimization was handled by a stochastic gradient descent (SGD) optimizer with a momentum of 0.9, a weight decay of 5×10^{-4}, and an initial learning rate of 10^{-3}. Additionally, we implemented a "poly" learning rate scheduling strategy [60] with a power factor of 0.9 to facilitate smooth convergence. All input images were standardized to a resolution of 320×320 pixels.

Table 5 presents a comprehensive quantitative comparison of various salient object detection (SOD) models across five benchmark datasets: *DUTS-TE*, *DUT-OMRON*, *HKU-IS*, *ECSSD*, and *PASCAL-S*. The evaluation utilizes several key metrics, including the maximum F-measure (F_{max}), average F-measure (F_{avg}), weighted F-measure (F_β^ω), E-measure (E_m), S-measure (S_m), and Mean Absolute Error (MAE). Performance results are categorized by the architectural backbones used, specifically *VGG-16* and the *ResNet* family (*ResNet-50*, *ResNet-101*, and *ResNeXt-101*). To highlight the leading methodologies, the best, second-best, and third-best results for each metric are color-coded in red, green, and blue, respectively. Across nearly all datasets and backbones, the proposed *MINet* consistently achieves top-tier performance, often securing the highest values in F-measure and S-measure while maintaining the lowest MAE, which validates its effectiveness in producing accurate and structurally coherent saliency maps compared to state-of-the-art methods like *EGNet*, *SCRN*, and *CPD*. In addition to the numerical results presented in Table 5, Figure 5 provides a visual quantitative comparison through Precision-Recall (PR) curves and F-measure curves across all five benchmark datasets. The solid red line representing our MINet consistently demonstrates superior performance and stability.

Figure 6 illustrates a qualitative comparison of visual saliency maps generated by various state-of-the-art salient object detection (SOD) models. The figure is structured as a grid where the leftmost column displays the original input images, followed by the Ground Truth (GT) masks, and then the results from several competing frameworks, including *MINet*, *AFNet*, *PAGE*, *CPD*, *C2SNet*, *RAS*, *PAGR*, *PiCANet* [62], *DSS*, *UCF*, *MSRNet*, *NLDF*, and *Amulet*. Each row represents a different challenging scenario, such as identifying multiple small objects, capturing intricate details of a cat or bird, or distinguishing salient targets in cluttered and low-contrast backgrounds. Visually, the *MINet* results align most closely with the ground truth, effectively suppressing background noise and producing solid, well-defined foreground masks. In contrast, older or less robust methods like UCF, Amulet, and NLDF often exhibit fragmented predictions, blurred boundaries, or significant false positives in complex regions.

The qualitative results presented in the visual comparison highlight the robust feature extraction and noise suppression capabilities of the proposed *MINet* architecture. As observed in the provided saliency maps, our model consistently generates results that align most closely with the Ground

Table 5. The following results provide a numerical assessment of various saliency detection frameworks across five standard benchmark datasets. Performance is measured using the maximum and average Fβ (higher values indicate superior results), the Sm (higher is better), and the MAE (lower is better). To distinguish the leading methodologies, the top three performing models are highlighted in red, green, and blue, respectively. An asterisk (*) denotes the application of additional post-processing techniques. Furthermore, the abbreviations X101 and R101 refer to the implementation of ResNeXt-101 [24] and ResNet-101 [61] as the respective architectural backbones.

Model	DUTS-TE						DUT-OMRON						HKU-IS						ECSSD						Pascal-S					
	Fmax	*Favg*	$F\omega_\beta$	*Em*	*Sm*	MAE	*Fmax*	*Favg*	$F\omega_\beta$	*Em*	*Sm*	MAE	*Fmax*	*Favg*	$F\omega_\beta$	*Em*	*Sm*	MAE	*Fmax*	*Favg*	$F\omega_\beta$	*Em*	*Sm*	MAE	*Fmax*	*Favg*	$F\omega_\beta$	*Em*	*Sm*	MAE
VGG-16																														
Ours	0.877	0.823	0.813	0.912	0.875	0.039	0.794	0.741	0.719	0.864	0.822	0.057	0.932	0.906	0.892	0.955	0.914	0.030	0.943	0.922	0.905	0.947	0.919	0.036	0.882	0.843	0.820	0.898	0.855	0.065
EGNet19	0.877	0.800	0.797	0.895	0.878	0.044	0.809	0.744	0.728	0.864	0.836	0.057	0.927	0.893	0.875	0.950	0.910	0.035	0.943	0.913	0.892	0.941	0.919	0.041	0.871	0.821	0.798	0.873	0.848	0.078
AFNet19	0.863	0.793	0.785	0.895	0.867	0.046	0.797	0.739	0.717	0.860	0.826	0.057	0.925	0.889	0.872	0.949	0.906	0.036	0.935	0.908	0.886	0.942	0.914	0.042	0.871	0.828	0.804	0.887	0.850	0.071
MLMSNet19	0.852	0.745	0.761	0.863	0.862	0.049	0.774	0.692	0.681	0.839	0.809	0.064	0.920	0.871	0.860	0.938	0.907	0.039	0.928	0.868	0.871	0.916	0.911	0.045	0.864	0.771	0.785	0.847	0.845	0.075
PAGE19	0.838	0.777	0.769	0.886	0.854	0.052	0.792	0.736	0.722	0.860	0.825	0.062	0.920	0.884	0.868	0.948	0.904	0.036	0.931	0.906	0.886	0.943	0.912	0.042	0.859	0.817	0.792	0.879	0.840	0.078
HRS19	0.843	0.793	0.746	0.889	0.829	0.051	0.762	0.708	0.645	0.842	0.772	0.066	0.913	0.892	0.854	0.938	0.883	0.042	0.920	0.902	0.859	0.923	0.883	0.054	0.852	0.809	0.748	0.850	0.801	0.090
CPD19	0.864	0.813	0.801	0.908	0.867	0.043	0.794	0.745	0.715	0.868	0.818	0.057	0.924	0.896	0.881	0.952	0.904	0.033	0.936	0.914	0.894	0.943	0.910	0.040	0.873	0.832	0.806	0.884	0.843	0.074
C2SNet18	0.811	0.717	0.717	0.847	0.831	0.062	0.759	0.682	0.663	0.828	0.799	0.072	0.898	0.851	0.834	0.928	0.886	0.047	0.911	0.865	0.854	0.915	0.896	0.053	0.857	0.775	0.777	0.850	0.840	0.080
RAS18	0.831	0.751	0.740	0.864	0.839	0.059	0.787	0.713	0.695	0.849	0.814	0.062	0.913	0.871	0.843	0.931	0.887	0.045	0.921	0.889	0.857	0.922	0.893	0.056	0.838	0.787	0.738	0.837	0.795	0.104
PAGR18	0.854	0.784	0.724	0.883	0.838	0.055	0.771	0.711	0.622	0.843	0.775	0.071	0.919	0.887	0.823	0.941	0.889	0.047	0.927	0.894	0.834	0.917	0.889	0.061	0.858	0.808	0.738	0.854	0.817	0.093
PiCANet18	0.851	0.749	0.747	0.865	0.861	0.054	0.794	0.710	0.691	0.842	0.826	0.068	0.922	0.870	0.848	0.938	0.905	0.042	0.931	0.885	0.865	0.926	0.914	0.046	0.871	0.804	0.781	0.862	0.851	0.077
DSS* 17	0.825	0.789	0.755	0.885	0.824	0.056	0.781	0.740	0.697	0.844	0.790	0.063	0.916	0.902	0.867	0.935	0.878	0.040	0.899	0.863	0.822	0.907	0.873	0.068	0.843	0.812	0.762	0.848	0.795	0.096
UCF17	0.773	0.631	0.596	0.770	0.782	0.112	0.730	0.621	0.574	0.768	0.760	0.120	0.888	0.823	0.780	0.904	0.874	0.061	0.903	0.844	0.806	0.896	0.884	0.069	0.825	0.738	0.700	0.809	0.807	0.115
MSRNet17	0.829	0.723	0.720	0.848	0.839	0.061	0.782	0.687	0.670	0.827	0.808	0.073	0.914	0.866	0.853	0.940	0.903	0.040	0.911	0.868	0.850	0.918	0.895	0.054	0.858	0.790	0.769	0.854	0.841	0.081
NLDF17	0.812	0.739	0.710	0.855	0.816	0.065	0.753	0.684	0.634	0.817	0.770	0.080	0.902	0.872	0.839	0.929	0.878	0.048	0.905	0.878	0.839	0.912	0.875	0.063	0.833	0.782	0.742	0.842	0.804	0.099
AMU17	0.778	0.678	0.658	0.803	0.804	0.085	0.743	0.647	0.626	0.784	0.781	0.098	0.899	0.843	0.819	0.915	0.886	0.050	0.915	0.868	0.840	0.912	0.894	0.059	0.841	0.771	0.741	0.831	0.821	0.098
ResNet-50/ResNet-101/ResNeXt-101																														
Ours-R	0.884	0.828	0.825	0.917	0.884	0.037	0.810	0.756	0.738	0.873	0.833	0.055	0.935	0.908	0.899	0.961	0.920	0.028	0.947	0.924	0.911	0.953	0.925	0.033	0.882	0.842	0.821	0.899	0.857	0.064
SCRN19	0.888	0.809	0.803	0.901	0.885	0.040	0.811	0.746	0.720	0.869	0.837	0.056	0.935	0.897	0.878	0.954	0.917	0.033	0.950	0.918	0.899	0.942	0.927	0.037	0.890	0.839	0.816	0.888	0.867	0.065
EGNet-R19	0.889	0.815	0.816	0.907	0.887	0.039	0.815	0.756	0.738	0.874	0.841	0.053	0.935	0.901	0.887	0.956	0.918	0.031	0.947	0.920	0.903	0.947	0.925	0.037	0.878	0.831	0.807	0.879	0.853	0.075
CPD-R19	0.865	0.805	0.795	0.904	0.869	0.043	0.797	0.747	0.719	0.873	0.825	0.056	0.925	0.891	0.876	0.952	0.906	0.034	0.939	0.917	0.898	0.949	0.918	0.037	0.872	0.831	0.803	0.887	0.847	0.072
ICNet19	0.855	0.767	0.762	0.880	0.865	0.048	0.813	0.739	0.730	0.859	0.837	0.061	0.925	0.880	0.858	0.943	0.908	0.037	0.938	0.880	0.881	0.923	0.918	0.041	0.866	0.786	0.790	0.860	0.850	0.071
BANet19	0.872	0.815	0.811	0.907	0.879	0.040	0.803	0.746	0.736	0.865	0.832	0.059	0.930	0.899	0.887	0.955	0.913	0.032	0.945	0.923	0.908	0.953	0.924	0.035	0.879	0.838	0.817	0.889	0.853	0.070
BASNet19	0.859	0.791	0.803	0.884	0.866	0.048	0.805	0.756	0.751	0.869	0.836	0.056	0.930	0.898	0.890	0.947	0.908	0.033	0.942	0.879	0.904	0.921	0.916	0.037	0.863	0.781	0.800	0.853	0.837	0.077
CapSal^{R101} 19	0.823	0.755	0.691	0.866	0.815	0.062	0.639	0.564	0.484	0.703	0.674	0.096	0.884	0.843	0.782	0.907	0.850	0.058	0.862	0.825	0.771	0.866	0.826	0.074	0.869	0.827	0.791	0.878	0.837	0.074
R3Net*X101 18	0.833	0.787	0.767	0.879	0.836	0.057	0.795	0.748	0.728	0.859	0.817	0.062	0.915	0.894	0.878	0.945	0.895	0.035	0.934	0.914	0.902	0.940	0.910	0.040	0.846	0.805	0.765	0.846	0.805	0.094
DGRL18	0.828	0.794	0.774	0.899	0.842	0.050	0.779	0.709	0.697	0.850	0.810	0.063	0.914	0.882	0.865	0.947	0.896	0.038	0.925	0.903	0.883	0.943	0.906	0.043	0.860	0.814	0.792	0.881	0.839	0.075
PiCANet-R18	0.860	0.759	0.755	0.873	0.869	0.051	0.803	0.717	0.695	0.848	0.832	0.065	0.919	0.870	0.842	0.941	0.905	0.044	0.935	0.886	0.867	0.927	0.917	0.046	0.870	0.804	0.782	0.862	0.854	0.076
SRM17	0.826	0.753	0.722	0.867	0.836	0.059	0.769	0.707	0.658	0.843	0.798	0.069	0.906	0.873	0.835	0.939	0.887	0.046	0.917	0.892	0.853	0.928	0.895	0.054	0.850	0.804	0.762	0.861	0.833	0.085

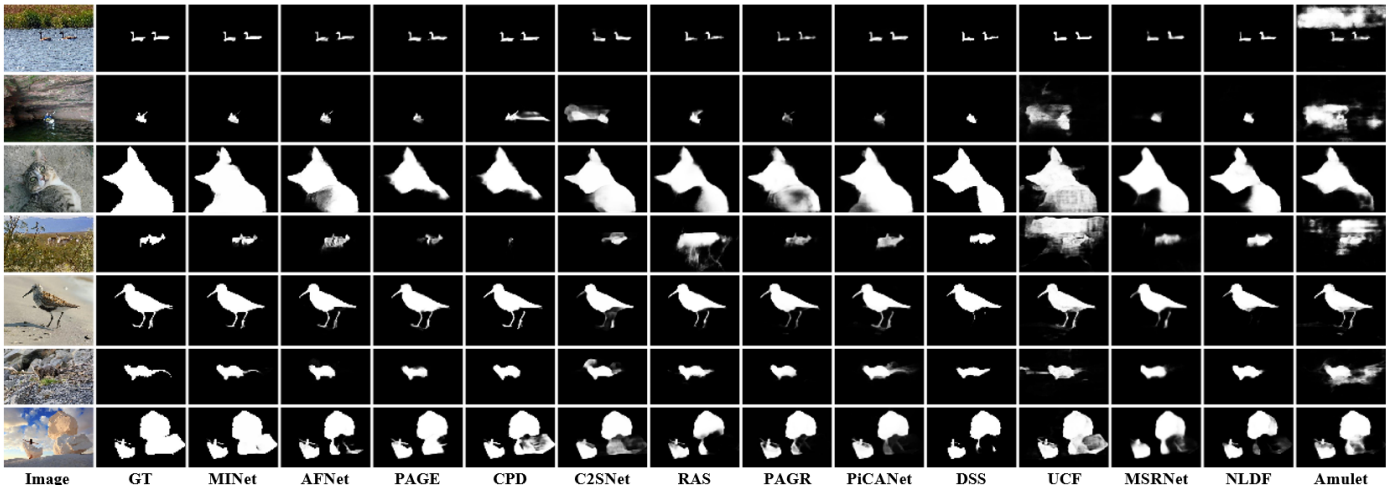

Figure 6. Comparative analysis of visuals in different methods.

Table 6. Comparative analysis of MINet: Previous Work vs. Proposed MINet.

Category / Approach	Key Methodology & Innovation	Major Advantages	Limitations & Trade-offs
Traditional Deep SOD Models	Employs standard Fully Convolutional Networks (FCNs) for pixel-wise classification.	Transitioned SOD from handcrafted priors to automated semantic learning.	Often produces blurry boundaries due to repeated pooling; lacks global structural context.
Context-Aware Modules (ASPP, PPM)	Utilizes Atrous Spatial Pyramid Pooling [12] or Pyramid Pooling Modules [13] at the deepest encoder layers.	Captures wide receptive fields to perceive objects of varying sizes.	Placement at top layers results in lost fine-grained spatial details due to low resolution.
Refinement-Based Frameworks	Uses secondary networks or recurrent refinement loops (e.g., CPD [8], AFNet [7]) to clean initial maps.	Effectively sharpens object boundaries and removes background artifacts.	High computational overhead and training complexity; often too slow for real-time use.
Mutual Learning [25]	Introduces collaborative interaction between peer networks to improve feature extraction.	Facilitates information exchange to enhance discriminative power.	Traditionally limited to classification tasks; requires adaptation for dense pixel-level prediction.
Recent Hybrid Models (2024-2025)	Integrates Transformer-based global reasoning (e.g., TAMF) or diffusion-based denoising (SOD-diffusion).	Achieves high adaptation to multi-scale objects and generates high-quality masks from noisy distributions.	May require significant computational resources for training foundational generative models.
Proposed Solution (MINet)	Introduces Aggregate Interaction (AIM) and Self-Interaction (SIM) modules.	Achieves state-of-the-art accuracy with 86.1 FPS inference speed; no post-processing required.	Performance is heavily dependent on the quality and diversity of the labeled training datasets.

Truth (*GT*), effectively managing scenarios involving both intricate fine structures and heavy background clutter. In cases featuring thin appendages or complex textures, such as the legs of the bird or the fur of the cat, *MINet* maintains structural integrity where competing methods like *UCF*, Amulet, and *NLDF* often produce fragmented or blurred outputs. Furthermore, in environments with low foreground-background contrast or distracting natural elements, the integration of the Aggregate Interaction Module (AIM) and Self-Interaction Module (*SIM*) allows the network to filter out non-salient regions that cause significant false positives in models like *MSRNet* and *DSS*. By leveraging the Consistency-Enhanced Loss (*CEL*), *MINet* ensures that the interior of salient objects remains uniformly highlighted and "solid," providing a level of spatial coherence that eliminates the need for traditional post-processing techniques. A comprehensive comparison between the proposed MINet and prior methodologies across key architectural and functional dimensions is provided in Table 6.

The proposed framework demonstrates superior computational efficiency and real-time viability

Table 7. Comparative analysis of MINet: Previous Work vs. Proposed MINet.

Method	Backbone	Parameters (M)	FLOPs (G)	FPS (GPU)	No Post-processing
Proposed MINet	VGG-16	162.38	87.1	86.1	Yes
Proposed MINet	ResNet-50	47.52	24.5	50.2	Yes
GCPANet (2020)	ResNet-50	67.06	32.5	81.3	Yes
F3Net (2020)	ResNet-50	25.54	12.8	24.2	Yes
EGNet (2019)	VGG-16	108.07	156.0	11.2	Yes
CPD (2019)	ResNet-50	29.23	14.5	68.1	Yes
PoolNet (2019)	VGG-16	68.26	108.3	31.2	Yes
BASNet (2019)	ResNet-34	70.32	127.4	25.3	No (CRF)
SCRN (2019)	ResNet-50	25.24	18.2	34.2	Yes

as shown in Table 7, achieving a high inference speed of 86.1 FPS through the architecturally streamlined Aggregate Interaction (AIM) and Self-Interaction (SIM) modules. Unlike traditional models that rely on heavy multi-scale pyramid plugins or recurrent refinement loops, these modules utilize minimal sampling rates to reduce redundant feature processing, resulting in a significantly optimized accuracy-to-FLOPs ratio (all FLOPs are calculated based on a standard input resolution of 320 × 320). Furthermore, the integration of the Consistency-Enhanced Loss (CEL) eliminates the need for memory-intensive post-processing tools like Conditional Random Fields (CRF), ensuring a cleaner end-to-end deployment pipeline with a reduced memory footprint and lower latency. This modular efficiency is further validated by the ResNet-50 variant, which maintains competitive performance with only 47.52 M parameters, proving that the proposed interaction mechanisms offer a robust and scalable solution for high-performance salient object detection without the typical computational overhead.

5 Conclusion and Future Scope

In this research, *MINet*, a novel multi-scale interactive network, is presented to address the persistent challenges of scale variation and background noise in salient object detection. By integrating the Aggregate Interaction Module (*AIM*) and the Self-Interaction Module (*SIM*), the architecture effectively facilitates a reciprocal exchange of information between different resolutions and internal feature branches. This mutual learning strategy allows the model to preserve fine structural details while simultaneously capturing global semantic context. Furthermore, the introduction of the Consistency-Enhanced Loss (*CEL*) ensures spatial uniformity within salient regions, significantly reducing fragmentation without the need for computationally expensive post-processing. Extensive quantitative and qualitative evaluations across five benchmark datasets like *DUTS*, *DUT-OMRON*, *ECSSD*, *HKU-IS*, and *PASCAL-S* demonstrate that *MINet* consistently outperforms state-of-the-art methods.

With a high inference speed of 86.1 FPS, the proposed model achieves an optimal balance between accuracy and real-time efficiency, making it highly suitable for practical computer vision applications.

While *MINet* demonstrates robust performance in static image saliency, several avenues remain for future exploration. To further advance this research, we propose the following directions:

a. **Video Salient Object Detection**: Extending the interactive modules to incorporate temporal dynamics for consistent object tracking and saliency estimation in video sequences.

b. **Cross-Domain Generalization**: Investigating the model's adaptability to specialized domains such as medical imaging (e.g., polyp or tumor segmentation) and high-resolution aerial surveillance.

c. **Lightweight Optimization**: Further refining the architecture for deployment on edge devices and mobile platforms by exploring depth-wise separable convolutions or knowledge distillation techniques.

d. **Weakly Supervised Learning**: Adapting the multi-scale interaction framework to work with scribbles or image-level labels to reduce the dependency on labor-intensive pixel-level ground truth annotations.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The author declares no conflicts of interest.

AI Use Statement

The author declares that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Mahadevan, V., & Vasconcelos, N. (2009, June). Saliency-based discriminant tracking. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 1007-1013). IEEE. [CrossRef]

[2] Gao, Y., Wang, M., Tao, D., Ji, R., & Dai, Q. (2012). 3-D object retrieval and recognition with hypergraph analysis. *IEEE transactions on image processing, 21*(9), 4290-4303. [CrossRef]

[3] Rosin, P. L., & Lai, Y. K. (2013). Artistic minimal rendering with lines and blocks. *Graphical Models, 75*(4), 208-229. [CrossRef]

[4] Wang, T., Piao, Y., Lu, H., Li, X., & Zhang, L. (2019, October). Deep Learning for Light Field Saliency Detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 8837-8847). IEEE. [CrossRef]

[5] Wang, X., Liang, X., Yang, B., & Li, F. W. (2019). No-reference synthetic image quality assessment with convolutional neural network and local image saliency. *Computational visual media, 5*(2), 193-208. [CrossRef]

[6] Achanta, R., Hemami, S., Estrada, F., & Susstrunk, S. (2009, June). Frequency-tuned salient region detection. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 1597-1604). IEEE. [CrossRef]

[7] Feng, M., Lu, H., & Ding, E. (2019, June). Attentive Feedback Network for Boundary-Aware Salient Object Detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1623-1632). IEEE. [CrossRef]

[8] Wu, R., Feng, M., Guan, W., Wang, D., Lu, H., & Ding, E. (2019, June). A Mutual Learning Method for Salient Object Detection With Intertwined Multi-Supervision. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 8142-8151). IEEE. [CrossRef]

[9] Zhang, P., Wang, D., Lu, H., Wang, H., & Ruan, X. (2017, October). Amulet: Aggregating Multi-level Convolutional Features for Salient Object Detection. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 202-211). IEEE. [CrossRef]

[10] Yu, S., Zhang, B., Xiao, J., & Lim, E. G. (2021, May). Structure-consistent weakly supervised salient object detection with local saliency coherence. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 35, No. 4, pp. 3234-3242). [CrossRef]

[11] Borji, A., Cheng, M. M., Hou, Q., Jiang, H., & Li, J. (2019). Salient object detection: A survey. *Computational visual media, 5*(2), 117-150. [CrossRef]

[12] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence, 40*(4), 834-848. [CrossRef]

[13] Zhang, P., Wang, D., Lu, H., Wang, H., & Yin, B. (2017, October). Learning Uncertain Convolutional Features for Accurate Saliency Detection. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 212-221). IEEE. [CrossRef]

[14] Deng, Z., Hu, X., Zhu, L., Xu, X., Qin, J., Han, G., & Heng, P. A. (2018, July). R3net: recurrent residual refinement network for saliency detection. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence* (pp. 684-690).

[15] Wang, T., Borji, A., Zhang, L., Zhang, P., & Lu, H. (2017, October). A Stagewise Refinement Model for Detecting Salient Objects in Images. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 4039-4048). IEEE. [CrossRef]

[16] Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., & Jagersand, M. (2019, June). BASNet: Boundary-Aware Salient Object Detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 7471-7481). IEEE. [CrossRef]

[17] Li, G., & Yu, Y. (2015). Visual saliency based on multiscale deep features. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5455-5463).[CrossRef]

[18] Shelhamer, E., Long, J., & Darrell, T. (2017). Fully Convolutional Networks for Semantic Segmentation. *IEEE transactions on pattern analysis and machine intelligence, 39*(4), 640-651. [CrossRef]

[19] Zhao, R., Ouyang, W., Li, H., & Wang, X. (2015, June). Saliency detection by multi-context deep learning. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1265-1274). IEEE. [CrossRef]

[20] Margolin, R., Zelnik-Manor, L., & Tal, A. (2014, June). How to Evaluate Foreground Maps. In *2014 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248-255). IEEE. [CrossRef]

[21] Li, G., Xie, Y., Lin, L., & Yu, Y. (2017). Instance-Level

Salient Object Segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 247-256). IEEE. [CrossRef]

[22] Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017, July). Pyramid Scene Parsing Network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6230-6239). IEEE. [CrossRef]

[23] Wu, Z., Su, L., & Huang, Q. (2019, June). Cascaded Partial Decoder for Fast and Accurate Salient Object Detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3902-3911). IEEE. [CrossRef]

[24] Wu, Z., Su, L., & Huang, Q. (2019, October). Stacked Cross Refinement Network for Edge-Aware Salient Object Detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 7263-7272). IEEE. [CrossRef]

[25] Zhang, Y., Xiang, T., Hospedales, T. M., & Lu, H. (2018, June). Deep Mutual Learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4320-4328). IEEE. [CrossRef]

[26] Wei, G., Zhou, M., Sun, J., Shi, X., Yin, M., Zhao, X., & Lin, X. (2026). Robust salient object detection based on triple attention-guided multi-resolution fusion and feature refinement. *PLoS One, 21*(2), e0342974. [CrossRef]

[27] Li, X., Yang, F., Cheng, H., Liu, W., & Shen, D. (2018, September). Contour Knowledge Transfer for Salient Object Detection. In *European Conference on Computer Vision* (pp. 370-385). Cham: Springer International Publishing. [CrossRef]

[28] Xu, Y., Li, X., Yuan, H., Yang, Y., & Zhang, L. (2023). Multi-task learning with multi-query transformer for dense prediction. *IEEE Transactions on Circuits and Systems for Video Technology, 34*(2), 1228-1240. [CrossRef]

[29] Zheng, Z., & Peng, Y. (2025). Boundary-Aware Cross-Level Multi-Scale Fusion Network for RGB-D Salient Object Detection. *IEEE Access, 13*, 48271-48285. [CrossRef]

[30] Zhao, J., Wen, X., He, Y., Yang, X., & Song, K. (2024). Wavelet-Driven Multi-Band Feature Fusion for RGB-T Salient Object Detection. *Sensors, 24*(24), 8159. [CrossRef]

[31] Ge, Y., Pan, J., Ren, J., He, M., Bi, H., & Zhang, Q. (2025). Co-salient object detection with consensus mining and consistency cross-layer interactive decoding. *Image and Vision Computing, 154*, 105414. [CrossRef]

[32] Zhang, S., Huang, J., Chen, S., Wu, Y., Hu, T., & Liu, J. (2024, October). SOD-diffusion: Salient Object Detection via Diffusion-Based Image Generators. In *Computer Graphics Forum* (Vol. 43, No. 7, p. e15251). [CrossRef]

[33] Aldubaikhi, A., & Patel, S. (2025). Advancements in small-object detection (2023–2025): Approaches, datasets, benchmarks, applications, and practical guidance. *Applied Sciences, 15*(22), 11882. [CrossRef]

[34] Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017, July). Aggregated Residual Transformations for Deep Neural Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5987-5995). IEEE. [CrossRef]

[35] Yan, Q., Xu, L., Shi, J., & Jia, J. (2013, June). Hierarchical Saliency Detection. In *2013 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1155-1162). IEEE. [CrossRef]

[36] Guo, F., Wang, W., Shen, J., Shao, L., Yang, J., Tao, D., & Tang, Y. Y. (2017). Video saliency detection using object proposals. *IEEE transactions on cybernetics, 48*(11), 3159-3170. [CrossRef]

[37] Zhang, L., Ai, J., Jiang, B., Lu, H., & Li, X. (2017). Saliency detection via absorbing Markov chain with learnt transition probability. *IEEE Transactions on image processing, 27*(2), 987-998. [CrossRef]

[38] Hou, Q., Cheng, M. M., Hu, X., Borji, A., Tu, Z., & Torr, P. H. S. (2018). Deeply Supervised Salient Object Detection with Short Connections. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(4), 815-828. [CrossRef]

[39] Wang, T., Zhang, L., Wang, S., Lu, H., Yang, G., Ruan, X., & Borji, A. (2018, June). Detect Globally, Refine Locally: A Novel Approach to Saliency Detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3127-3135). IEEE. [CrossRef]

[40] Wang, W., Lai, Q., Fu, H., Shen, J., Ling, H., & Yang, R. (2021). Salient object detection in the deep learning era: An in-depth survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(6), 3239-3259. [CrossRef]

[41] Wu, R., Feng, M., Guan, W., Wang, D., Lu, H., & Ding, E. (2019). A mutual learning method for salient object detection with intertwined multi-supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8150-8159). [CrossRef]

[42] Zhang, L., Yang, C., Lu, H., Ruan, X., & Yang, M. H. (2016). Ranking saliency. *IEEE transactions on pattern analysis and machine intelligence, 39*(9), 1892-1904. [CrossRef]

[43] Liu, N., Zhang, N., & Han, J. (2020). Learning selective self-mutual attention for RGB-D saliency detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13756-13765). [CrossRef]

[44] Pang, Y., Zhao, X., Zhang, L., & Lu, H. (2020). Multi-scale interactive network for salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9413-9422). [CrossRef]

[45] Chen, S., Tan, X., Wang, B., & Hu, X. (2018, September). Reverse Attention for Salient Object Detection. In

European Conference on Computer Vision (pp. 236-252). Cham: Springer International Publishing. [CrossRef]

[46] Yang, C., Zhang, L., Lu, H., Ruan, X., & Yang, M. H. (2013, June). Saliency Detection via Graph-Based Manifold Ranking. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3166-3173). [CrossRef]

[47] Perazzi, F., Krähenbühl, P., Pritch, Y., & Hornung, A. (2012, June). Saliency filters: Contrast based filtering for salient region detection. In *2012 IEEE conference on computer vision and pattern recognition* (pp. 733-740). IEEE. [CrossRef]

[48] Krähenbühl, P., & Koltun, V. (2011, December). Efficient inference in fully connected CRFs with Gaussian edge potentials. In *Proceedings of the 25th International Conference on Neural Information Processing Systems* (pp. 109-117).

[49] Zhao, J., Liu, J. J., Fan, D. P., Cao, Y., Yang, J., & Cheng, M. M. (2019, October). EGNet: Edge Guidance Network for Salient Object Detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 8778-8787). IEEE. [CrossRef]

[50] Li, Y., Hou, X., Koch, C., Rehg, J. M., & Yuille, A. L. (2014). The Secrets of Salient Object Segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 280-287). IEEE. [CrossRef]

[51] Cheng, M. M., Mitra, N. J., Huang, X., Torr, P. H., & Hu, S. M. (2014). Global contrast based salient region detection. *IEEE transactions on pattern analysis and machine intelligence, 37*(3), 569-582. [CrossRef]

[52] Su, J., Li, J., Zhang, Y., Xia, C., & Tian, Y. (2019, October). Selectivity or Invariance: Boundary-Aware Salient Object Detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 3798-3807). IEEE. [CrossRef]

[53] Zeng, Y., Zhang, P., Lin, Z., Zhang, J., & Lu, H. (2019, October). Towards High-Resolution Salient Object Detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 7233-7242). IEEE. [CrossRef]

[54] Wei, Y., Wen, F., Zhu, W., & Sun, J. (2012, October). Geodesic saliency using background priors. In *European conference on computer vision* (pp. 29-42). Berlin, Heidelberg: Springer Berlin Heidelberg. [CrossRef]

[55] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017, July). Feature Pyramid Networks for Object Detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 936-944). IEEE. [CrossRef]

[56] Zhang, L., Dai, J., Lu, H., He, Y., & Wang, G. (2018, June). A Bi-Directional Message Passing Model for Salient Object Detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1741-1750). IEEE. [CrossRef]

[57] Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., & Ruan, X. (2017, July). Learning to Detect Salient Objects with Image-Level Supervision. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3796-3805). IEEE. [CrossRef]

[58] Fan, D. P., Cheng, M. M., Liu, Y., Li, T., & Borji, A. (2017, October). Structure-Measure: A New Way to Evaluate Foreground Maps. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 4558-4567). IEEE. [CrossRef]

[59] Fan, D. P., Gong, C., Cao, Y., Ren, B., Cheng, M. M., & Borji, A. (2018, July). Enhanced-alignment measure for binary foreground map evaluation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence* (pp. 698-704).

[60] Liu, W., Rabinovich, A., & Berg, A. C. (2015). Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*.

[61] He, K., Zhang, X., Ren, S., & Sun, J. (2016, June). Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770-778). IEEE. [CrossRef]

[62] Liu, N., Han, J., & Yang, M. H. (2018, June). PiCANet: Learning Pixel-Wise Contextual Attention for Saliency Detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3089-3098). IEEE. [CrossRef]

[63] Luo, Z., Mishra, A., Achkar, A., Eichel, J., Li, S., & Jodoin, P. M. (2017, July). Non-local Deep Features for Salient Object Detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6593-6601). IEEE Computer Society. [CrossRef]

[64] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[65] Wang, W., Shen, J., Cheng, M. M., & Shao, L. (2019, June). An Iterative and Cooperative Top-Down and Bottom-Up Inference Network for Salient Object Detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5961-5970). IEEE. [CrossRef]

[66] Wang, W., Zhao, S., Shen, J., Hoi, S. C., & Borji, A. (2019, June). Salient Object Detection With Pyramid Attention and Salient Edges. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1448-1457). IEEE. [CrossRef]

[67] Zhang, L., Zhang, J., Lin, Z., Lu, H., & He, Y. (2019, June). CapSal: Leveraging Captioning to Boost Semantics for Salient Object Detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6017-6026). IEEE. [CrossRef]

[68] Zhang, X., Wang, T., Qi, J., Lu, H., & Wang, G. (2018, June). Progressive Attention Guided Recurrent Network for Salient Object Detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 714-722). IEEE. [CrossRef]