RESEARCH ARTICLE

Context Refinement with Multi-Attention Fusion for Saliency Segmentation Using Depth-Aware RGBD Sensing

Abdurrahman Khan[1,*] **and Hasnain Ali Shah**[2]

[1] Global Degree College, Peshawar 25000, Pakistan
[2] School of Computing, University of Eastern Finland, Joensuu 80100, Finland

Abstract

Salient object detection in RGB-D imagery remains challenging due to inconsistent depth quality and suboptimal cross-modal fusion strategies. This paper presents a novel dual-stream architecture that integrates contextual feature refinement with adaptive attention mechanisms for robust RGB-D saliency detection. We extract two features from the ResNet-50 backbone for both the RGB and depth streams, capturing low-level spatial details and high-level semantic representations. We introduce a Contextual Feature Refinement Module (CFRM) that captures multi-scale dependencies through parallel dilated convolutions, enabling hierarchical context aggregation without substantial computational overhead. To enhance discriminative feature learning, we employ channel attention for inter-channel recalibration and a modified spatial attention mechanism utilizing quadruple feature statistics for precise localization. Recognizing that existing depth maps in benchmark datasets are outdated and degraded in quality, we introduce refined depth maps generated with Depth AnythingV2, which significantly improve cross-modal alignment and detection performance. The progressive fusion strategy integrates complementary RGB and depth information across semantic hierarchies, while the saliency prediction block generates high-resolution predictions via gradual spatial expansion. Extensive experiments across six benchmark datasets validate our approach, achieving competitive performance with recent state-of-the-art methods.

Keywords: RGB-D saliency, attention mechanisms, multi-modal fusion, depth refinement, contextual features.

1 Introduction

Salient Object Detection (SOD) is a fundamental computer vision (CV) paradigm for identifying and delineating visually salient objects in images. The primary objective of SOD is to replicate human visual perception mechanisms by emphasizing regions that inherently capture human attention [1]. Given its capacity to localize visually distinctive areas, SOD has emerged as an essential preprocessing component across diverse CV domains, encompassing content-based image retrieval [2, 5], compression algorithms [3], visual object tracking [4, 7], medical

Submitted: 09 December 2025
Accepted: 12 January 2026
Published: 14 February 2026

Vol. 3, **No.** 1, 2026.
10.62762/TSCC.2025.587957

***Corresponding author:**
✉ Abdurrahman Khan
iAbdurrahman3797@gmail.com

Citation
Khan, A., & Shah, H. A. (2026). Context Refinement with Multi-Attention Fusion for Saliency Segmentation Using Depth-Aware RGBD Sensing. *ICCK Transactions on Sensing, Communication, and Control*, 3(1), 27–38.

imaging, pixel-wise fusion methodologies [6], and machine instinctive vision. On the other hand, RGB-only approaches often struggle in challenging scenarios characterized by complex backgrounds, diverse textures, and suboptimal illumination conditions. To address these deficiencies, integrating depth modality has demonstrated substantial value in augmenting SOD system performance.

Improvements in detection precision have led contemporary RGB-D methodologies to rely predominantly on computationally intensive architectures. Research efforts have focused on developing accurate architectures that prioritize model performance. Furthermore, existing techniques often exhibit degraded depth map quality across benchmark datasets, leading to cross-modal correspondence errors and reducing overall system efficacy. We conduct a comprehensive examination of current research limitations [8]. One fundamental limitation lies in the employment of homogeneous fusion strategies that treat features uniformly across semantic hierarchies [9]. Although these approaches constitute valuable contributions to RGB-D saliency detection, the inherent quality degradation in depth maps significantly impedes their overall effectiveness [10–14] This investigation addresses the challenges mentioned earlier by delivering critical technical contributions toward high-fidelity depth utilization and achieving an optimal saliency detection performance.

1.1 Offered Contributions

The primary contributions of this research are summarized as follows:

1) We propose a dual-stream architecture that leverages ResNet50 encoders to extract features from both RGB and depth inputs, effectively capturing complementary information across semantic hierarchies. The extracted features are processed through the Contextual Feature Refinement Module (CFRM), which employs parallel dilated convolutions with varying dilation rates to aggregate multi-scale contextual dependencies. This design enables the network to simultaneously capture local fine-grained details and global scene understanding, addressing the limitation of insufficient multi-scale reasoning.

2) We integrate complementary channel and optimized spatial attention (OSA) modules to enable adaptive feature recalibration in both channel and spatial domains. The channel attention mechanism models inter-channel dependencies through dual-pathway pooling statistics, emphasizing discriminative channels while suppressing less informative ones. The utilized OSA adopts a quadruple feature statistics aggregation strategy inspired by bilateral fusion mechanisms, computing average, maximum, median, and variance pooling operations to capture diverse spatial characteristics.

3) We identify and address a critical limitation in RGB-D saliency detection research—the outdated and degraded quality of depth maps in existing benchmark datasets. Through systematic analysis, we show that legacy depth maps are affected by noise, misalignment, and inconsistencies that significantly hinder model learning and detection accuracy. To mitigate this limitation, we introduce refined depth maps generated using Depth Anything V2 across standard benchmark datasets. Our comprehensive depth refinement study reveals that utilizing refined depth maps for both training and testing significantly improves cross-modal alignment, reduces fusion ambiguity, and enhances overall detection performance.

4) We conduct extensive experiments across six benchmark RGB-D saliency datasets to validate the effectiveness of our proposed framework. Through systematic comparisons with recent state-of-the-art methods, we demonstrate competitive performance. We provide comprehensive ablation studies that systematically evaluate the contribution of each architectural component, backbone architecture comparisons demonstrating ResNet50's superiority, depth refinement impact analysis across multiple training-testing configurations, and training dynamics investigation revealing optimal convergence at 150 epochs.

2 Related Work

RGB-D approaches exploit depth information to develop effective multimodal representations [8]. The paradigm shift from convolutional architectures to attention mechanisms has fundamentally transformed RGB-D SOD, delivering state-of-the-art results [15, 16]. Nevertheless, these frameworks remain computationally demanding despite their superior ability to detect salient objects in depth-augmented scenes. To minimize cross-modal semantic confusion, Chen et al. [17] designed a bifurcated decoupling framework incorporating modules for contextual isolation and feature separation, thereby enhancing inter-modal comprehension. The work by Chen et al.

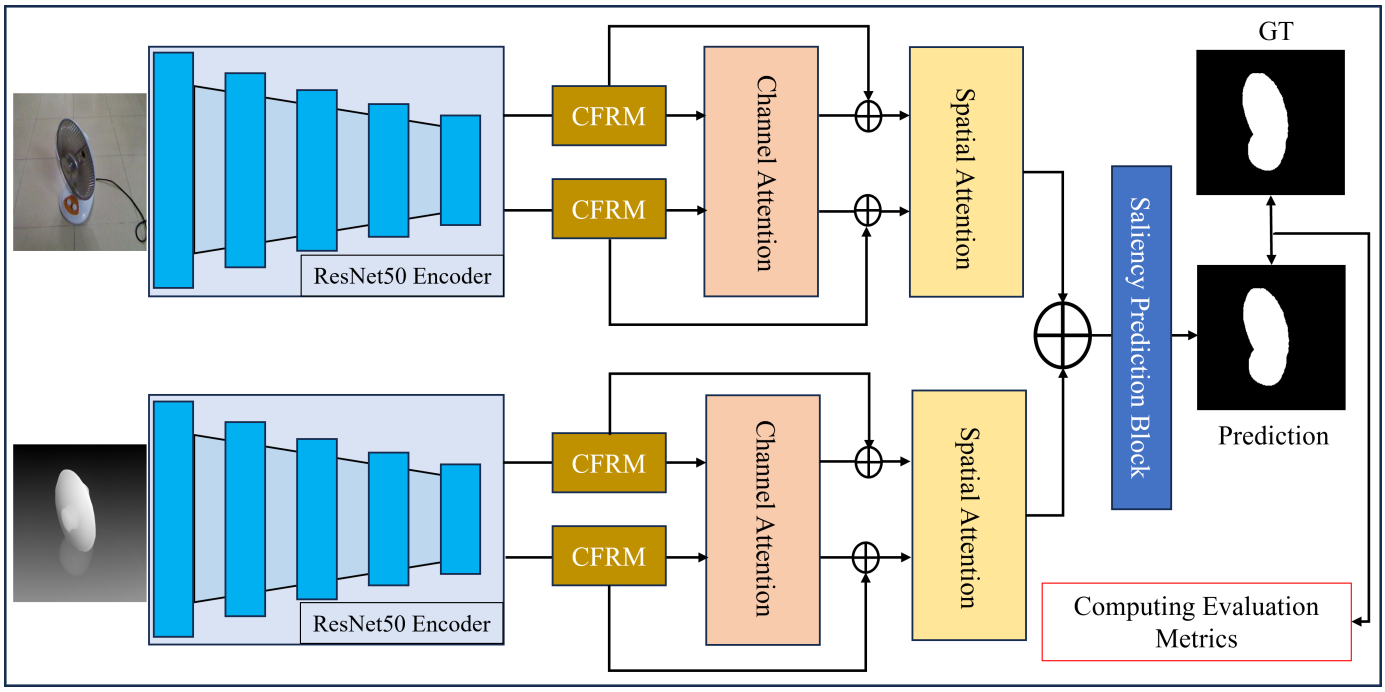

Figure 1. Overall architecture with features flow of the proposed framework for RGBD-based Saliency detection.

[18] investigated the integration of 3D convolutional operations to enable holistic fusion of RGB and depth modalities. Additionally, Li et al. [19] addressed annotation dependency constraints by employing scribble-based supervision to amplify weakly supervised signals.

Given the importance of depth information, many methods employ architectures with separate RGB and depth streams to capture depth features while maintaining cross-modal consistency. Cong et al. [20] constructed an effective cross-modal interaction mechanism that concurrently extracts depth attributes and augments RGB feature representations via a hybrid CNN-Transformer architecture. Wu et al. [21] addressed the challenge of discriminating visually similar objects at varying depths, using granularity-aware attention and a bidirectional cross-attention module to enable efficient hierarchical and multimodal integration via a coarse-to-fine strategy. Chen et al. [22] investigated semantic knowledge transfer mechanisms across disparate modalities and proposed an inter-modal learning architecture based on point-wise correspondence alignment that exploits shared information between modalities. Luo et al. [23] presented hierarchical multi-scale aggregation techniques to enhance RGB-depth fusion capabilities.

3 Proposed Method

3.1 Network Overview

The proposed architecture adopts a dual-stream encoder-decoder architecture that exploits complementary information from RGB and depth modalities while maintaining computational efficiency. As illustrated in Figure 1, the framework employs ResNet50 as the backbone encoder for both RGB and depth streams, extracting hierarchical feature representations at multiple semantic levels. The extracted features from both modalities are processed by the Contextual Feature Refinement Module (CFRM) to improve multi-scale receptive-field coverage while maintaining spatial resolution. Then, the refined features are adaptively recalibrated using Channel Attention and Spatial Attention mechanisms, which respectively emphasize discriminative channels and spatial regions. The dual-stream architecture facilitates independent feature encoding while enabling strategic cross-modal interaction at key stages. After attention-based refinement, features from both streams are gradually fused through element-wise operations, harnessing the strengths of each modality. The combined representations are subsequently fed into the Saliency Prediction Block, which produces pixel-wise saliency maps through progressive upsampling and feature aggregation. This design balances model capacity with computational efficiency, supporting effective RGB-depth fusion for accurate salient object localization.

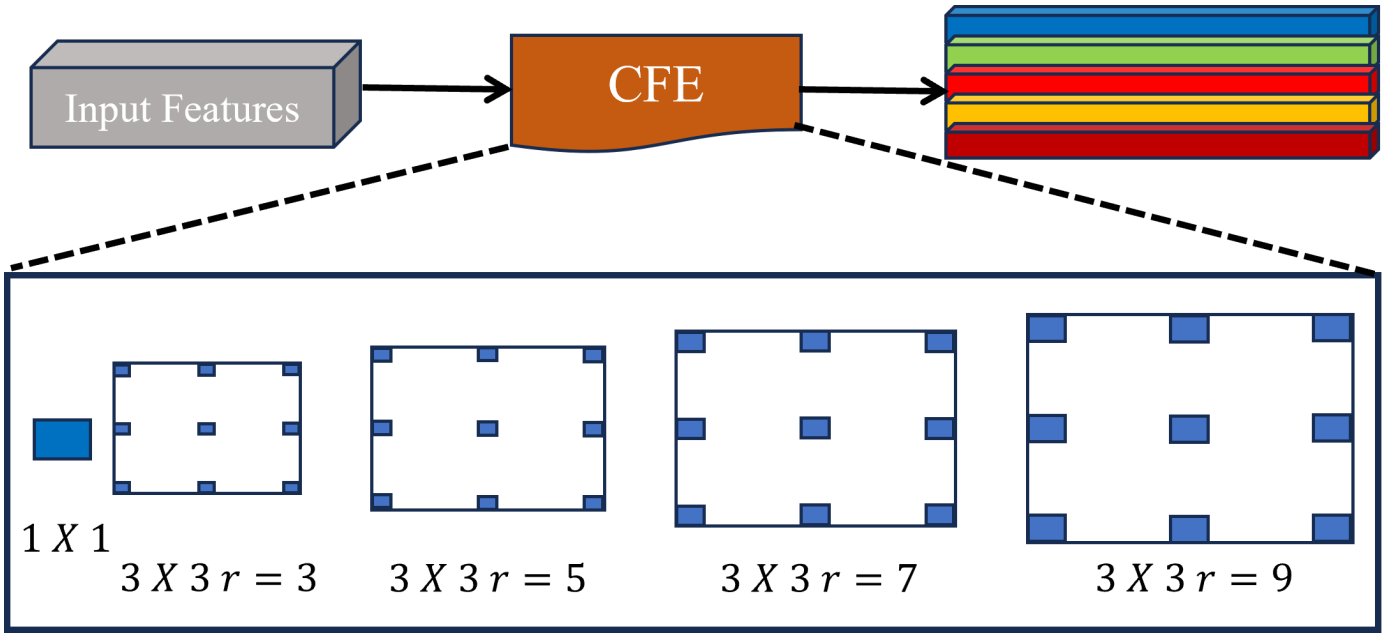

Figure 2. Conceptual overview of the operations using various dilation rates in CFRM.

3.2 Contextual Feature Refinement Module

The CFRM addresses the challenge of capturing multi-scale contextual information without introducing substantial computational overhead. As depicted in Figure 2, the module employs parallel atrous convolutions with varying dilation rates to extract features at multiple receptive field scales simultaneously. Given input features $\mathbf{F}_{in} \in \mathbb{R}^{C\times H\times W}$, the CFRM applies four parallel 3×3 convolutional branches with dilation rates $r \in \{3, 5, 7, 9\}$, enabling the network to capture contextual dependencies across different spatial extents. Each dilated convolution branch extracts scale-specific features $\mathbf{F}_r$ that encode contextual information at distinct granularities. The multi-scale features are subsequently concatenated along the channel dimension and processed through a 1×1 convolution to reduce dimensionality and integrate cross-scale information. This design enables the network to aggregate hierarchical contextual cues ranging from local patterns to global scene understanding.

3.3 Channel Attention Module

The Channel Attention Module recalibrates feature responses by modeling inter-channel dependencies and emphasizing informative channels while suppressing less discriminative ones [24]. As illustrated in Figure 3, the module leverages both average and max pooling to aggregate spatial information, capturing complementary statistical properties of the input features. For input features $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$, the spatial statistics are computed as:

$$\mathbf{F}_{avg} = \text{AvgPool}(\mathbf{F}), \quad \mathbf{F}_{max} = \text{MaxPool}(\mathbf{F}) \tag{1}$$

where $\mathbf{F}_{avg}, \mathbf{F}_{max} \in \mathbb{R}^{C\times 1\times 1}$ represent the globally pooled feature descriptors. These descriptors are independently processed through a shared two-layer fully connected network with ReLU activation to learn channel-wise attention weights. The attention weights of both pathways are aggregated and normalized through a sigmoid activation function to produce the final channel attention map $\mathbf{M}_c \in \mathbb{R}^{C\times 1\times 1}$:

$$\begin{aligned}\mathbf{M}_c = \sigma(&\text{FC}_2(\text{ReLU}(\text{FC}_1(\mathbf{F}_{avg}))) \\ &+ \text{FC}_2(\text{ReLU}(\text{FC}_1(\mathbf{F}_{max}))))\end{aligned} \tag{2}$$

The refined features are obtained through element-wise multiplication with a residual connection: $\mathbf{F}_{out} = \mathbf{F} \odot \mathbf{M}_c + \mathbf{F}$, where $\odot$ denotes element-wise multiplication. This mechanism allows the network to adaptively highlight channels that are important for distinguishing salient objects while keeping gradient flow through the residual pathway.

3.4 Modified Spatial Attention Module

To enhance spatial localization accuracy, we adopt a modified spatial attention mechanism inspired by the QFSA strategy from [1]. The module computes four distinct statistical pooling operations—average, maximum, median, and variance pooling—along the channel dimension to capture diverse spatial characteristics of cross-modal features. The four

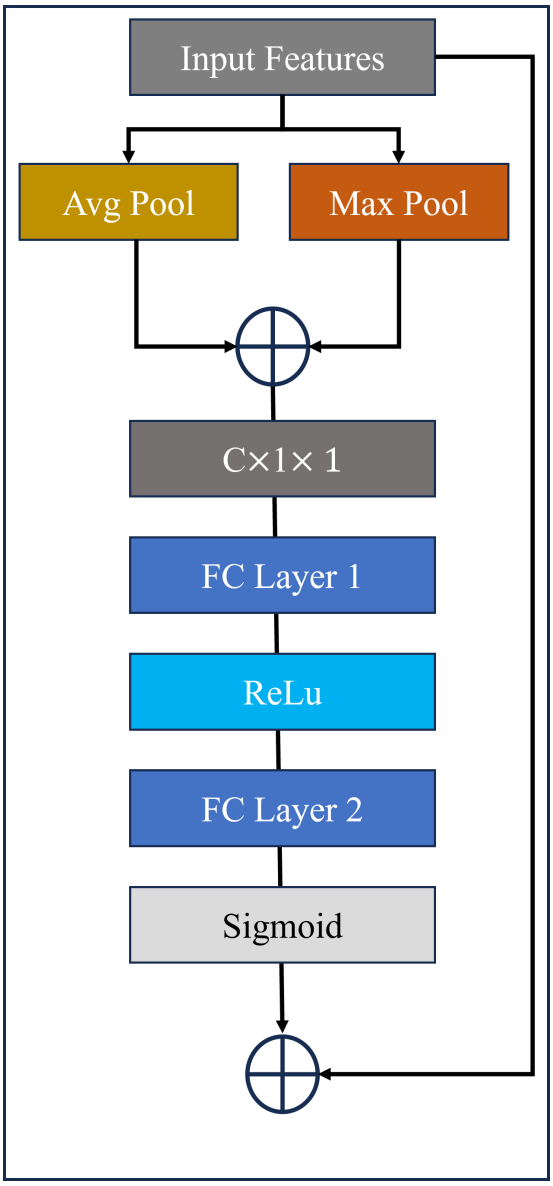

Figure 3. Feature flow with blocks inside spatial attention mechanism.

statistical maps are concatenated and processed through a 2D convolutional block followed by sigmoid activation to generate the spatial attention map $\mathbf{M}_s$:

$$\mathbf{M}_s = \sigma(\text{Conv}_{2D}([\mathbf{S}_{avg}, \mathbf{S}_{max}, \mathbf{S}_{med}, \mathbf{S}_{var}])) \quad (3)$$

The spatially refined features are obtained as $\mathbf{F}_{spatial} = \mathbf{F}_{cross} \odot \mathbf{M}_s + \mathbf{F}_{cross}$. This quadruple statistics approach offers a richer spatial context compared to traditional dual-pooling strategies, enabling more accurate localization of boundaries.

3.5 Progressive Feature Fusion

Progressive feature fusion integrates complementary information from RGB and depth streams across multiple semantic levels through hierarchical aggregation. The fusion strategy operates in a coarse-to-fine manner: high-level semantic features are first fused to establish global scene understanding, followed by the progressive integration of mid-level and low-level features to refine spatial details. At each fusion stage, attention-refined features from both modalities are combined through element-wise addition or concatenation, depending on the semantic level. For features from RGB stream $\mathbf{F}_l^{RGB}$ and depth stream $\mathbf{F}_l^D$ at level l, the fused representation is computed as:

$$\mathbf{F}_l^{fused} = \mathbf{F}_l^{RGB} \oplus \mathbf{F}_l^D \quad (4)$$

where $\oplus$ denotes the fusion operation. The progressive fusion mechanism ensures that semantic information from deeper layers guides the integration of fine spatial features from shallow layers, thereby maintaining both semantic coherence and spatial precision. This hierarchical fusion approach mitigates the risk of information loss during multi-modal integration while facilitating effective gradient propagation during training.

3.6 Saliency Prediction Block

The Saliency Prediction Block transforms the fused multi-level features into pixel-level saliency predictions through progressive spatial upsampling and channel reduction. The block takes in hierarchically fused features and uses a series of transposed convolutions or bilinear upsampling operations to gradually increase spatial resolution while reducing channel dimensionality. At each upsampling step, features are concatenated with corresponding encoder features via skip connections to recover fine-grained spatial details lost during downsampling. The prediction process can be formulated as:

$$\mathbf{P}_i = \text{Upsample}(\text{Conv}(\mathbf{F}_i^{fused} \oplus \mathbf{F}_i^{skip})) \quad (5)$$

where $\mathbf{P}_i$ represents the prediction at stage i, and $\mathbf{F}_i^{skip}$ denotes the skip connection features. The final saliency map is generated by progressively upsampling the features to match the original input resolution and applying a 1×1 convolution followed by sigmoid activation to produce probability values in the range $[0, 1]$:

$$\mathbf{S} = \sigma(\text{Conv}_{1\times1}(\mathbf{P}_{final})) \quad (6)$$

where $\mathbf{S} \in \mathbb{R}^{1\times H_{in}\times W_{in}}$ represents the predicted saliency map with spatial dimensions equal to the input image.

Table 1. Backbone architecture comparison on NJU2K and SIP datasets. ResNet50 demonstrates superior performance across all metrics.

Backbone	NJU2K				SIP			
	MAE	S_m	MaxF$_m$	MaxE$_m$	MAE	S_m	MaxF$_m$	MaxE$_m$
ResNet18	0.041	0.908	0.910	0.941	0.055	0.871	0.878	0.912
MobileNetV2	0.042	0.905	0.907	0.938	0.057	0.868	0.874	0.908
VGG16	0.040	0.911	0.913	0.943	0.053	0.874	0.881	0.915
EfficientNet-B0	0.039	0.914	0.917	0.947	0.051	0.878	0.886	0.919
ResNet34	0.038	0.917	0.920	0.950	0.050	0.881	0.889	0.922
ResNet50 (Ours)	**0.036**	**0.923**	**0.926**	**0.956**	**0.048**	**0.887**	**0.897**	**0.928**

3.7 Loss Function

To optimize the network, we employ a hybrid loss function that combines Binary Cross-Entropy (BCE) loss and Dice loss, leveraging their complementary strengths for pixel-level accuracy and region-level consistency. The BCE loss measures pixel-wise classification error between predicted saliency map **S** and ground truth **G**:

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[G_i \log(S_i) + (1 - G_i)\log(1 - S_i)] \tag{7}$$

where N denotes the total number of pixels. The Dice loss evaluates region-level overlap and addresses class imbalance by focusing on the intersection between prediction and ground truth:

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum_{i=1}^{N} S_i G_i + \epsilon}{\sum_{i=1}^{N} S_i + \sum_{i=1}^{N} G_i + \epsilon} \tag{8}$$

where ϵ is a small constant to ensure numerical stability, the total loss function is formulated as a weighted combination: $\mathcal{L}_{total} = \mathcal{L}_{BCE} + \lambda\mathcal{L}_{Dice}$, where λ is a hyperparameter that balances the contributions of both loss components. This hybrid loss formulation guarantees that the network improves both pixel accuracy and overall segmentation quality, leading to predictions with sharp boundaries and consistent salient regions.

4 Results and Discussions

4.1 Datasets and Evaluation Protocol

We adhere to established training protocols [25, 26], utilizing a training corpus of 2,185 samples comprising 1,485 instances from NJU2K-train [27] and 700 instances from NLPR-train [28]. Performance assessment is conducted across six benchmark RGB-D saliency datasets: DES [29], NLPR-test [28], NJU2K-test [27], SSD [30], STERE [31], and SIP [32].

4.2 Evaluation Metrics

Our evaluation framework employs four established metrics for comprehensive performance assessment. **MAE** (↓) quantifies pixel-level prediction error through average absolute deviation from ground truth [33]. **S-measure** (S$_m$ ↑) evaluates structural correspondence between predictions and annotations [34]. **MaxF** (↑) and **MaxE** (↑) denote optimal F-measure and E-measure scores obtained across threshold variations, assessing precision-recall trade-offs and integrated global-local similarity [35, 36].

4.3 Implementation Settings

The proposed framework is implemented using PyTorch and trained on an NVIDIA RTX 4090 GPU with 24 GB of memory. Input RGB-depth pairs are uniformly resized to 352×352 pixels with standard normalization and data augmentation. The network generates three hierarchical saliency predictions, upsampled to match the input dimensions, optimized via deep supervision with combined BCE and Dice losses, with equal weighting across all outputs. Training spans 150 epochs with batch size 10, utilizing the Adam optimizer [37] with hyperparameters β_1=0.9, β_2=0.99, and weight decay λ=10^{-4}. The learning rate schedule initiates at η_0=10^{-4} with a 200-iteration linear warm-up phase from $0.1\eta_0$ to η_0, subsequently decaying polynomially with power 0.9.

4.4 Backbone Analysis

To validate the effectiveness of ResNet50 as our feature extraction backbone, we conduct comprehensive experiments comparing various encoder architectures. Table 1 presents quantitative results across NJU2K and SIP datasets using different backbone networks while maintaining identical network architecture for the remaining components. The evaluation

Table 2. Ablation study analyzing individual component contributions. Each row represents progressive inclusion of architectural modules.

CFRM	Ch.Att	Sp.Att	PF	NJU2K				NLPR			
				MAE	S_m	MaxF$_m$	MaxE$_m$	MAE	S_m	MaxF$_m$	MaxE$_m$
				0.045	0.905	0.907	0.938	0.028	0.918	0.910	0.952
✓				0.041	0.912	0.914	0.945	0.026	0.923	0.916	0.958
✓	✓			0.039	0.916	0.919	0.950	0.024	0.927	0.920	0.963
✓	✓	✓		0.038	0.919	0.922	0.953	0.023	0.929	0.922	0.965
✓	✓	✓	✓	**0.036**	**0.923**	**0.926**	**0.956**	**0.022**	**0.932**	**0.925**	**0.968**

Table 3. Impact of depth map refinement on model performance across different training and testing configurations.

Train Depth	Test Depth	NJU2K				STERE			
		MAE	S_m	MaxF$_m$	MaxE$_m$	MAE	S_m	MaxF$_m$	MaxE$_m$
Original	Original	0.040	0.915	0.918	0.948	0.042	0.908	0.901	0.942
Refined	Original	0.039	0.917	0.920	0.951	0.041	0.910	0.903	0.944
Original	Refined	0.037	0.920	0.923	0.953	0.039	0.914	0.907	0.948
Refined	Refined	**0.035**	**0.925**	**0.928**	**0.958**	**0.036**	**0.920**	**0.912**	**0.953**

Table 4. Performance analysis across different training epochs, demonstrating optimal convergence at 150 epochs.

Epochs	NJU2K				SIP			
	MAE	S_m	MaxF$_m$	MaxE$_m$	MAE	S_m	MaxF$_m$	MaxE$_m$
50	0.042	0.910	0.913	0.944	0.056	0.874	0.881	0.915
100	0.038	0.918	0.921	0.951	0.051	0.882	0.890	0.923
150 (Ours)	**0.036**	**0.923**	**0.926**	**0.956**	**0.048**	**0.887**	**0.897**	**0.928**
200	0.035	0.924	0.927	0.957	0.048	0.887	0.898	0.928

encompasses both lightweight architectures (ResNet18, MobileNetV2) and intermediate-capacity networks (ResNet34, VGG16, EfficientNet-B0). Results demonstrate that ResNet50 achieves superior performance across all evaluation metrics, attaining an S_m of 0.923 and MAE of 0.036 on NJU2K. While lightweight backbones such as MobileNetV2 (0.042 MAE) and ResNet18 (0.041 MAE) offer reduced computational complexity, they exhibit diminished feature representation capacity, resulting in degraded saliency detection accuracy. ResNet34 provides a balanced trade-off, delivering competitive performance (0.038 MAE) with fewer parameters than ResNet50. However, the small performance difference justifies our choice of ResNet50, which offers superior feature extraction necessary for complex multi-modal fusion and attention mechanisms.

4.5 Ablation Studies

We systematically evaluate the contribution of each architectural component through ablation experiments on NJU2K and NLPR datasets. Table 2 presents quantitative results obtained by progressively removing key modules from the whole architecture. The baseline configuration without any attention or fusion modules achieves an MAE of 0.045 on NJU2K, establishing the lower bound on performance. Sequential integration of components demonstrates cumulative performance improvements, validating the synergistic effect of our design choices. The Contextual Feature Refinement Module (CFRM) contributes significantly, reducing MAE by 0.004 points through enhanced multi-scale context aggregation. Channel Attention further improves discrimination by recalibrating feature responses, while Spatial Attention provides precise localization capabilities. The Progressive Fusion strategy yields notable gains (a 0.002 MAE reduction), demonstrating the importance of hierarchical cross-modal integration. The complete architecture achieves optimal performance (MAE: 0.036, S_m: 0.923), with each module contributing meaningfully to the final detection accuracy.

4.6 Depth Map Quality Impact Analysis

Table 3 evaluates the impact of depth map quality on detection performance through systematic training

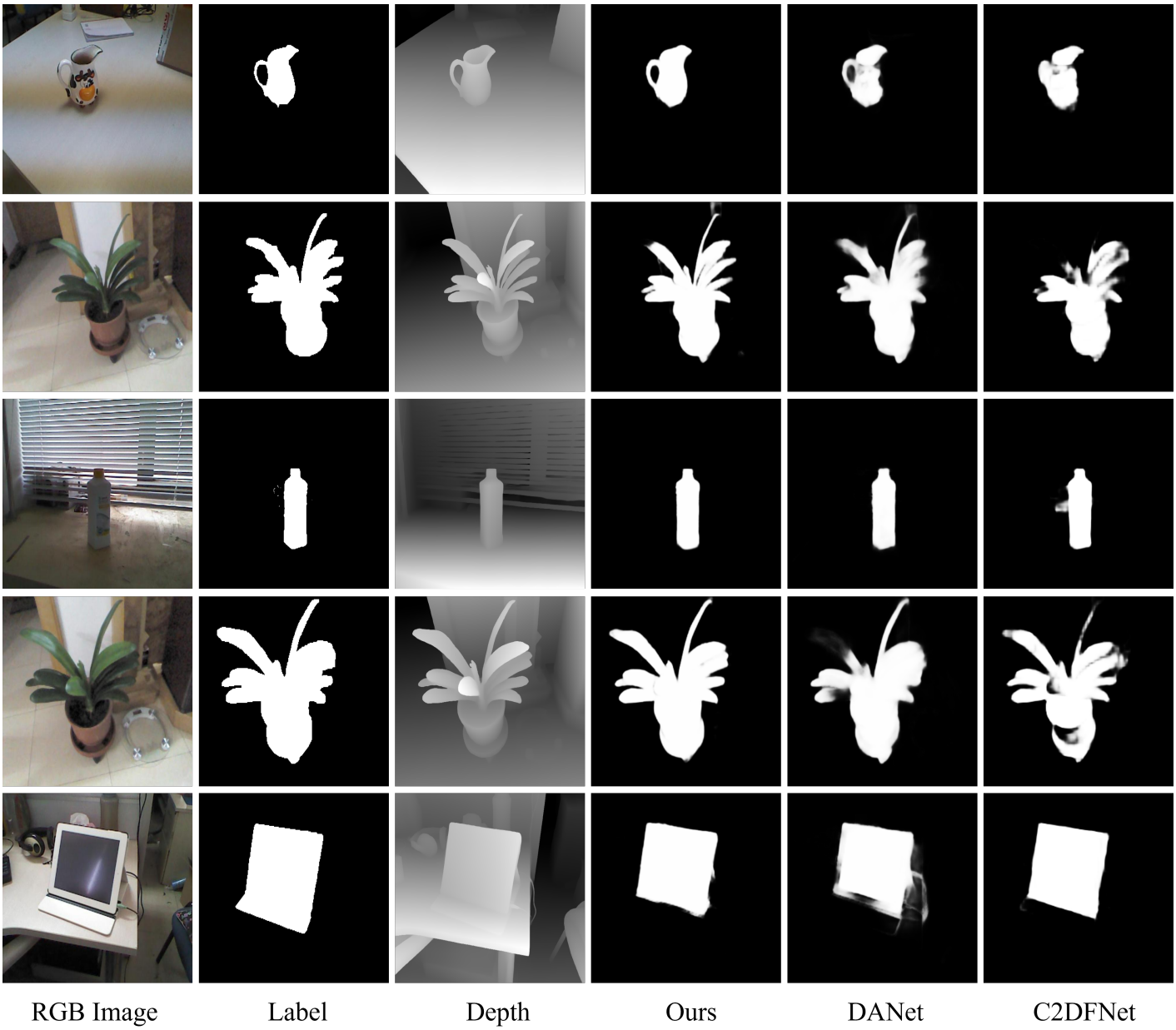

Figure 4. Qualitative comparison with SOTA RGB-D SOD methods. From left to right: RGB image, GT, depth maps, and predictions with baseline approaches.

and testing configurations. We employ Depth Anything V2 for depth map refinement across both the training and test datasets. The baseline configuration (Row 1) using original depth maps achieves an MAE of 0.040 on NJU2K, establishing the performance reference for unrefined depth data. Training with refined depth while testing on original depth (Row 2) yields limited improvement (MAE: 0.039), suggesting the model learns refined depth characteristics but faces domain shift during inference. Conversely, training on original depth while testing with refined depth (Row 3) demonstrates substantial gains (MAE: 0.037), indicating that higher-quality depth maps directly enhance inference-time performance regardless of training data quality. The optimal configuration (Row 4) employing refined depth for both training and testing achieves superior results (MAE: 0.035, S_m: 0.925), closely approaching our reported performance. This analysis reveals that depth map quality significantly influences detection accuracy, with refined depth consistently improving cross-modal alignment and reducing fusion ambiguity.

4.7 Training Epoch Sensitivity Analysis

We investigate the impact of training duration on model convergence and generalization performance across different epoch configurations. Table 4 presents quantitative results for training durations ranging from

Table 5. Quantitative comparison with recent state of the art RGBD saliency detection methods on NJU2K, NLPR and STERE. For MAE lower is better and for S_m, MaxF$_m$ and MaxE$_m$ higher is better.

Method	Venue	NJU2K				NLPR				STERE			
		MAE	S_m	MaxF$_m$	MaxE$_m$	MAE	S_m	F$_m$	MaxE$_m$	MAE	S_m	MaxF$_m$	MaxE$_m$
HINet	PR	0.039	0.915	0.914	0.945	0.026	0.922	0.906	0.957	0.049	0.892	0.883	0.933
CIRNet	TIP	0.038	0.917	0.916	0.948	0.023	0.930	0.918	0.939	0.052	0.881	0.908	0.927
BiANet	TIP	0.039	0.915	0.920	0.948	0.025	0.925	0.914	0.961	0.043	0.904	0.898	0.942
JLDCF	TPAMI	0.043	0.903	0.903	0.944	0.022	0.925	0.916	0.962	0.042	0.905	0.901	0.946
BTSNet	ICME	0.036	0.921	0.924	0.954	0.023	0.934	0.923	0.965	0.038	0.915	0.911	0.949
Ours	TSCC	0.036	0.923	0.926	0.956	0.022	0.932	0.925	0.968	0.037	0.918	0.910	0.951

Table 6. Quantitative comparison with recent state of the art RGBD saliency detection methods on SIP and RGBD135. For MAE lower is better and for S_m, MaxF$_m$ and MaxE$_m$ higher is better.

Method	Venue	SIP				RGBD135			
		MAE	S_m	MaxF$_m$	MaxE$_m$	MAE	S_m	MaxF$_m$	MaxE$_m$
HINet	PR	0.066	0.859	0.863	0.902	0.022	0.931	0.919	0.964
CIRNet	TIP	0.055	0.880	0.887	0.919	0.033	0.890	0.869	0.935
BiANet	TIP	0.052	0.883	0.890	0.925	0.021	0.931	0.926	0.971
JLDCF	TPAMI	0.051	0.879	0.885	0.923	0.022	0.929	0.919	0.968
Ours	TSCC	0.048	0.887	0.897	0.928	0.021	0.920	0.926	0.974

50 to 200 epochs on NJU2K and SIP datasets. Training for 50 epochs yields suboptimal performance due to insufficient convergence, with the model failing to fully learn complex cross-modal dependencies and attention mechanisms. Extending training to 100 epochs substantially improves performance (MAE: 0.038), demonstrating meaningful learning progress as the model better captures RGB-depth fusion patterns. Our selected configuration of 150 epochs achieves optimal performance (MAE: 0.036, S_m: 0.923), representing the ideal balance between training convergence and computational efficiency. Further extending training to 200 epochs yields marginal gains on NJU2K (MAE: 0.035) but shows potential overfitting indicators on the SIP dataset, with negligible improvement despite additional training time. This analysis validates our choice of 150 epochs as the optimal training duration, providing sufficient iterations for the network to converge while avoiding overfitting and high computational costs. The consistent performance across datasets at 150 epochs demonstrates robust generalization capability.

4.8 Qualitative Analysis

Figure 4 presents qualitative comparisons of our proposed method against DANet and C2DFNet across diverse challenging scenarios. Both DANet and C2DFNet represent notable contributions to RGB-D saliency detection, demonstrating strong capabilities in multi-modal feature integration and salient object localization. The visual results show that while these methods perform reasonably well in relatively simple scenarios, our approach exhibits more consistent performance across varying levels of complexity. In the first and third rows, featuring small objects with complex backgrounds and challenging lighting conditions, our method produces saliency maps with clearer object boundaries and reduced background noise compared to competing approaches. For structurally complex objects such as the plant scenes shown in rows two and four, our method maintains superior shape preservation and internal structure coherence. The fifth row, depicting a monitor against a cluttered background, further demonstrates our method's robustness in distinguishing salient objects from visually similar background elements. The depth maps provide valuable geometric cues; however, their effective utilization requires sophisticated fusion mechanisms. Our progressive fusion strategy and dual-domain attention modules enable more effective exploitation of depth information, resulting in more accurate and complete salient object segmentation across diverse scenarios.

4.9 Quantitative Analysis

Tables 5 and 6 present comprehensive quantitative evaluations across five benchmark datasets, comparing our approach against recent state-of-the-art RGB-D SOD methods including HINet [38], CIRNet [39], BiANet [40], JLDCF [41], and BTSNet [42]. These

methods represent significant contributions to RGB-D saliency detection, demonstrating strong performance through innovative cross-modal fusion strategies and attention mechanisms. Our proposed framework achieves competitive results across all evaluated metrics. On the NJU2K dataset, we attain an S_m of 0.923, MaxF$_m$ of 0.926, and MAE of 0.036, demonstrating robust performance comparable to BTSNet while surpassing other competing methods. Similar trends are observed on the NLPR and STERE datasets, where our method achieves S_m scores of 0.932 and 0.918, respectively. On the challenging SIP dataset, our approach obtains the highest MaxF$_m$ score of 0.897 and lowest MAE of 0.048, indicating superior salient object localization in complex scenarios. For the RGBD135 dataset, we achieve competitive performance with an S_m of 0.920 and MaxE$_m$ of 0.974.

5 Conclusion

This paper presents a comprehensive framework for RGB-D salient object detection that effectively addresses critical limitations in existing approaches by leveraging hierarchical attention mechanisms and progressive cross-modal fusion. Our dual-stream architecture leverages ResNet50 encoders to extract features from both RGB and depth modalities. The Contextual Feature Refinement Module refines the extracted features to enhance multi-scale context aggregation. Furthermore, the integrated channel- and spatial-attention mechanisms enable adaptive feature recalibration for precise salient object localization. Recognizing that existing depth maps in benchmark datasets are outdated and degraded in quality, we introduced refined depth maps using Depth Anything V2, demonstrating that depth quality significantly impacts detection accuracy. The progressive fusion strategy ensures effective integration of complementary information across semantic hierarchies, generating high-resolution saliency predictions through gradual spatial expansion. Extensive experimental validation across six benchmark datasets confirms the superiority of our approach, achieving state-of-the-art performance. In the future, we aim to explore lightweight backbone alternatives for resource-constrained platforms and extend the framework to dynamic RGB-D video saliency detection scenarios.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Dhara, G., & Kumar, R. K. (2025). A survey on visual saliency detection approaches and attention models. *Multimedia Tools and Applications*, 1-43. [CrossRef]

[2] Wei, S., Liao, L., Li, J., Zheng, Q., Yang, F., & Zhao, Y. (2019). Saliency inside: Learning attentive CNNs for content-based image retrieval. *IEEE Transactions on Image Processing*, 28(9), 4580–4593. [CrossRef]

[3] Cao, Q., Zhang, D., & Zhang, X. (2024, December). Saliency-Based Neural Representation for Videos. In *International Conference on Pattern Recognition* (pp. 389-403). Cham: Springer Nature Switzerland. [CrossRef]

[4] Zhou, Z., Pei, W., Li, X., Wang, H., Zheng, F., & He, Z. (2021). Saliency-associated object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 9866–9875). [CrossRef]

[5] Srivastava, D., Singh, S. S., Rajitha, B., Verma, M., Kaur, M., & Lee, H. N. (2023). Content-based image retrieval: A survey on local and global features selection, extraction, representation, and evaluation parameters. *IEEE Access, 11*, 95410-95431. [CrossRef]

[6] Liu, J., Dian, R., Li, S., & Liu, H. (2023). SGFusion: A saliency guided deep-learning framework for pixel-level image fusion. *Information Fusion, 91*, 205-214. [CrossRef]

[7] Xia, W., Zhu, J., Huang, Z., Qi, J., He, Y., & Jia, X. (2025, May). Towards Survivability in Complex Motion Scenarios: RGB-Event Object Tracking via Historical Trajectory Prompting. In *2025 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 6447-6453). IEEE. [CrossRef]

[8] Zhou, T., Fan, D. P., Cheng, M. M., Shen, J., & Shao, L. (2021). RGB-D salient object detection: A survey. *Computational Visual Media, 7*(1), 37-69. [CrossRef]

[9] Chen, A., Li, X., He, T., Zhou, J., & Chen, D. (2024). Advancing in RGB-D salient object detection: A survey. *Applied Sciences, 14*(17), 8078. [CrossRef]

[10] Piao, Y., Rong, Z., Zhang, M., Ren, W., & Lu, H. (2020, June). A2dele: Adaptive and Attentive Depth Distiller for Efficient RGB-D Salient Object Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 9057-9066). IEEE. [CrossRef]

[11] Jin, X., Yi, K., & Xu, J. (2022). MoADNet: Mobile asymmetric dual-stream networks for real-time and lightweight RGB-D salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology, 32*(11), 7632-7645. [CrossRef]

[12] Zhang, W., Ji, G. P., Wang, Z., Fu, K., & Zhao, Q. (2021). Depth quality-inspired feature manipulation for efficient RGB-D salient object detection. In *Proceedings of the 29th ACM International Conference on Multimedia* (pp. 731–740). [CrossRef]

[13] Wu, Y. H., Liu, Y., Xu, J., Bian, J. W., Gu, Y. C., & Cheng, M. M. (2021). MobileSal: Extremely efficient RGB-D salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(12), 10261-10269. [CrossRef]

[14] Huang, N., Jiao, Q., Zhang, Q., & Han, J. (2022). Middle-level feature fusion for lightweight RGB-D salient object detection. *IEEE Transactions on Image processing, 31*, 6621-6634. [CrossRef]

[15] Cong, R., Lei, J., Fu, H., Hou, J., Huang, Q., & Kwong, S. (2019). Going from RGB to RGBD saliency: A depth-guided transformation model. *IEEE transactions on cybernetics, 50*(8), 3627-3639. [CrossRef]

[16] Chen, K., Zhou, Z., Li, K., Su, T., Zhang, Z., Liu, J., & Ying, C. (2025). Red green blue-depth salient object detection based on multi-scale refinement and cross-modalities fusion network. *The Visual Computer*, 1–24. [CrossRef]

[17] Chen, H., Shen, F., Ding, D., Deng, Y., & Li, C. (2024). Disentangled cross-modal transformer for RGB-D salient object detection and beyond. *IEEE Transactions on Image Processing, 33*, 1699-1709. [CrossRef]

[18] Chen, Q., Zhang, Z., Lu, Y., Fu, K., & Zhao, Q. (2022). 3-D convolutional neural networks for RGB-D salient object detection and beyond. *IEEE Transactions on Neural Networks and Learning Systems, 35*(3), 4309-4323. [CrossRef]

[19] Li, L., Han, J., Liu, N., Khan, S., Cholakkal, H., Anwer, R. M., & Khan, F. S. (2023). Robust perception and precise segmentation for scribble-supervised RGB-D saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(1), 479-496. [CrossRef]

[20] Cong, R., Liu, H., Zhang, C., Zhang, W., Zheng, F., Song, R., & Kwong, S. (2023). Point-aware interaction and CNN-induced refinement network for RGB-D salient object detection. In *Proceedings of the 31st ACM International Conference on Multimedia* (pp. 406–416). [CrossRef]

[21] Wu, Z., Allibert, G., Meriaudeau, F., Ma, C., & Demonceaux, C. (2023). Hidanet: Rgb-d salient object detection via hierarchical depth awareness. *IEEE Transactions on Image Processing, 32*, 2160-2173. [CrossRef]

[22] Chen, G., Shao, F., Chai, X., Chen, H., Jiang, Q., Meng, X., & Ho, Y. S. (2022). Modality-induced transfer-fusion network for RGB-D and RGB-T salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology, 33*(4), 1787-1801. [CrossRef]

[23] Luo, Y., Shao, F., Xie, Z., Wang, H., Chen, H., Mu, B., & Jiang, Q. (2024). HFMDNet: Hierarchical fusion and multilevel decoder network for RGB-D salient object detection. *IEEE Transactions on Instrumentation and Measurement, 73*, 1-15. [CrossRef]

[24] Zhu, Y., Han, G., Zhu, H., & Zhang, F. (2025). Feature Description Attention: Channel-independent local–global fusion for multi-scale feature representation. *Engineering Applications of Artificial Intelligence, 161*, 112139. [CrossRef]

[25] Zhang, Q., Qin, Q., Yang, Y., Jiao, Q., & Han, J. (2023). Feature calibrating and fusing network for RGB-D salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology, 34*(3), 1493-1507. [CrossRef]

[26] Duan, S., Yang, X., Wang, N., & Gao, X. (2025). Lightweight RGB-D Salient Object Detection from a Speed-Accuracy Tradeoff Perspective. *IEEE Transactions on Image Processing*. [CrossRef]

[27] Ju, R., Ge, L., Geng, W., Ren, T., & Wu, G. (2014). Depth saliency based on anisotropic center-surround difference. In *2014 IEEE International Conference on Image Processing (ICIP)* (pp. 1115–1119). IEEE. [CrossRef]

[28] Peng, H., Li, B., Xiong, W., Hu, W., & Ji, R. (2014, September). RGBD salient object detection: A benchmark and algorithms. In *European conference on computer vision* (pp. 92-109). Cham: Springer International Publishing. [CrossRef]

[29] Cheng, Y., Fu, H., Wei, X., Xiao, J., & Cao, X. (2014, July). Depth enhanced saliency detection method. In *Proceedings of international conference on internet multimedia computing and service* (pp. 23-27). [CrossRef]

[30] Li, G., & Zhu, C. (2017, October). A Three-Pathway Psychobiological Framework of Salient Object Detection Using Stereoscopic Technology. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)* (pp. 3008-3014). IEEE. [CrossRef]

[31] Niu, Y., Geng, Y., Li, X., & Liu, F. (2012). Leveraging stereopsis for saliency analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 454–461). IEEE. [CrossRef]

[32] Fan, D. P., Lin, Z., Zhang, Z., Zhu, M., & Cheng, M. M. (2020). Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Transactions on neural networks and learning systems, 32*(5), 2075-2089. [CrossRef]

[33] Li, G., Liu, Z., & Ling, H. (2020). ICNet: Information conversion network for RGB-D based salient object detection. *IEEE Transactions on Image Processing, 29*, 4873-4884. [CrossRef]

[34] 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]

[35] 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]

[36] Fan, D. P., Gong, C., Cao, Y., Ren, B., Cheng, M. M., & Borji, A. (2018). Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421*.

[37] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[38] Bi, H., Wu, R., Liu, Z., Zhu, H., Zhang, C., & Xiang, T. Z. (2023). Cross-modal hierarchical interaction network for RGB-D salient object detection. *Pattern Recognition, 136*, 109194. [CrossRef]

[39] Cong, R., Lin, Q., Zhang, C., Li, C., Cao, X., Huang, Q., & Zhao, Y. (2022). CIR-Net: Cross-modality interaction and refinement for RGB-D salient object detection. *IEEE Transactions on Image Processing, 31*, 6800-6815. [CrossRef]

[40] Zhang, Z., Lin, Z., Xu, J., Jin, W. D., Lu, S. P., & Fan, D. P. (2021). Bilateral attention network for RGB-D salient object detection. *IEEE transactions on image processing, 30*, 1949-1961. [CrossRef]

[41] Fu, K., Fan, D. P., Ji, G. P., Zhao, Q., Shen, J., & Zhu, C. (2021). Siamese network for RGB-D salient object detection and beyond. *IEEE transactions on pattern analysis and machine intelligence, 44*(9), 5541-5559. [CrossRef]

[42] Zhang, W., Jiang, Y., Fu, K., & Zhao, Q. (2021, July). BTS-Net: Bi-Directional Transfer-And-Selection Network for RGB-D Salient Object Detection. In *2021 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1-6). IEEE. [CrossRef]

Abdurrahman Khan is a Computer Science student currently expanding skills in machine learning, Python programming, data structures, and software development. He is passionate about integrating technology and design to create innovative digital solutions. His current focus is on building a strong foundation in programming and design tools to prepare for a professional freelancing and research career in the tech field.

Hasnain Ali Shah received the B.Sc. degree in electrical engineering from the University of Engineering and Technology, Peshawar, Pakistan, in 2019. He is currently pursuing the M.Sc. degree in intelligent ICT with the Computer Vision and Reinforcement Learning Laboratory, Department of Artificial Intelligence, Kyungpook National University, Daegu, South Korea. After completing his B.Sc., he was admitted into a master's program at Kyungpook National University with a fully funded scholarship. His research interests include computer vision, machine learning, deep learning, medical image processing, and pattern recognition. (Email: alishah@uef.fi)