RESEARCH ARTICLE

Learning Cross-Modal Collaboration via Pyramid Attention for RGB Thermal Sensing in Saliency Detection

Muhammad Zain Hassan[1], **Alexandros Gazis**[2,3], **Abdurrahman Khan**[4] **and Zainab Ghazanfar**[5,*]

[1] Department of Software Engineering, University of Haripur, Haripur 22620, Pakistan
[2] Democritus University of Thrace, Xanthi 67100, Greece
[3] Heriot-Watt University, Edinburgh EH14 4AS, United Kingdom
[4] Global Degree College, Peshawar 25000, Pakistan
[5] Department of AI and Software, Gachon University, Seongnam 13120, Republic of Korea

Abstract

RGB–thermal (RGB-T) salient object detection exploits complementary cues from visible and thermal sensors to maintain reliable performance in adverse environments. However, many existing methods (i) fuse modalities before sufficiently enhancing intra-modal semantics and (ii) are sensitive to modality discrepancies caused by heterogeneous sensor characteristics. To address these issues, we propose PACNet (Pyramid Attention Collaboration Network), a hierarchical RGB-T framework that jointly models multi-scale and global context and performs refinement-before-fusion with cross-modal collaboration. Specifically, Dense Atrous Spatial Pyramid Pooling (DASPP) captures multi-scale contextual cues across semantic stages, while Multi-Head Self-Attention (MHSA) establishes long-range dependencies for global context modeling. We further design a hierarchical feature integration scheme that constructs two complementary feature streams, preserving fine-grained spatial details and strengthening high-level semantics. These streams are refined using a cross-interactive dual-attention module that enables bidirectional interaction between spatial and channel attention, improving localization and semantic discrimination while mitigating modality imbalance. Experiments on three public benchmarks (VT821, VT1000, and VT5000) demonstrate that PACNet achieves state-of-the-art performance and delivers consistent gains in challenging conditions such as low illumination, thermal clutter, and multi-scale targets.

Keywords: salient object detection, RGB-thermal fusion, cross-interactive dual attention, multi-modal learning.

1 Introduction

Salient object detection (SOD) aims to identify and segment pixel-level objects or regions that capture

Submitted: 05 December 2025
Accepted: 29 December 2025
Published: 29 January 2026

Vol. 3, **No.** 1, 2026.
10.62762/TSCC.2025.210523

***Corresponding author:**
Zainab Ghazanfar
zainab@gachon.ac.kr

Citation
Hassan, M. Z, Gazis, A., Khan, A., & Ghazanfar, Z. (2026). Learning Cross-Modal Collaboration via Pyramid Attention for RGB Thermal Sensing in Saliency Detection. *ICCK Transactions on Sensing, Communication, and Control*, 3(1), 1–14.

human visual attention within an image, serving as a fundamental task in computer vision with applications spanning semantic segmentation [1], object tracking [2], and visual localization [3]. While deep learning has driven remarkable advances in RGB-based SOD [4], single-modality approaches lack robustness under challenging real-world conditions. RGB-based methods struggle significantly in adverse lighting scenarios such as low illumination, overexposure, and complex backgrounds [5], where degraded image quality severely compromises detection accuracy.

To address these limitations, researchers have explored supplementary modalities to compensate for RGB deficiencies. Initial efforts focused on RGB-Depth (RGB-D) modalities, where depth maps provide spatial and structural information [6]. However, depth sensors suffer from poor imaging quality under insufficient illumination and adverse weather conditions [28], limiting their practical deployment. As a promising alternative, thermal infrared sensors have emerged for RGB-Thermal (RGB-T) SOD tasks [8]. Thermal imaging captures radiation emitted by objects above absolute zero, remaining insensitive to lighting and weather conditions [9]. Unlike depth information, thermal modality effectively highlights object contours even in challenging environments [10], making the RGB-T combination ideal for robust SOD in complex scenes.

Despite recent progress, existing RGB-T SOD methods face several critical challenges. First, inherent modality differences arise from distinct sensor imaging properties, manifesting as varying sensitivities to scene interference and domain gaps in feature representations [12]. Current approaches typically adopt either single-flow paradigms that fuse multi-scale features during encoding [13], or dual-flow paradigms employing parallel decoders for independent modality processing [31]. However, these architectures often provide insufficient modality-specific supervision and struggle to handle defective inputs when one modality is severely degraded. Second, most existing methods follow a Cross-Modal then Cross-Scale paradigm: extracting multi-scale features, performing cross-modal fusion at corresponding scales, and integrating across scales using Feature Pyramid Network (FPN)-like decoders [15]. This approach neglects intra-modal semantic enhancement before fusion, failing to effectively express saliency instance information and compromising both fusion quality and generalization. Third, the timing and strategy for thermal information utilization remain under-explored. Unlike depth maps that directly relate to spatial perception, thermal images reflect temperature distributions and lack direct correlation with saliency [16]. There is no inherent assumption that hotter objects are more salient, and in some cases, salient objects may differ between RGB and thermal modalities, necessitating adaptive cross-modal integration strategies.

To address these challenges, we propose a novel RGB-T SOD framework that integrates hierarchical multi-scale features through strategic attention mechanisms and cross-modal collaboration. Rather than directly fusing raw multi-modal features, our approach first enhances intra-modal representations through Dense Atrous Spatial Pyramid Pooling (DASPP) modules that capture multi-scale contextual information at different semantic levels, while a Multi-Head Self-Attention (MHSA) mechanism establishes global contextual relationships. We then introduce a cross-interactive dual attention mechanism that processes two hierarchically integrated feature streams through spatial and channel attention pathways with bidirectional information exchange, enabling comprehensive feature refinement from complementary perspectives. This design ensures effective cross-modal collaboration while maintaining robustness to modality deficiencies and domain discrepancies.

The main contributions of this work are summarized as follows:

- **Hierarchical context integration (local-to-global)**: We propose a hierarchical feature integration network that jointly models multi-scale local context via DASPP and global dependencies via MHSA, producing two complementary representations: F_{123} (detail-preserving) and F_{345} (semantic-enriched).
- **Cross-interactive dual attention refinement**: We introduce a cross-interactive dual attention module that performs spatial attention and channel attention with bidirectional interaction, enabling coordinated refinement of localization cues and semantic discriminability.
- **Robust RGB–thermal fusion strategy**: We design a cross-modal fusion scheme that combines early additive fusion with attention-guided recalibration to reduce modality imbalance and improve robustness under challenging conditions (e.g., low illumination in RGB or noisy thermal

patterns).

- **Comprehensive validation on standard benchmarks**: Extensive experiments on VT821, VT1000, and VT5000 demonstrate state-of-the-art performance across multiple evaluation metrics, with consistent gains in challenging scenarios such as thermal clutter, low illumination, and multi-scale objects.

2 Related Work

2.1 RGB and RGB-D SOD

Traditional SOD methods primarily relied on hand-crafted features such as color contrast and edge density [17, 18], suffering from limited generalization capability. The emergence of deep learning revolutionized SOD through convolutional neural networks [19], enabling pixel-wise predictions with significantly improved accuracy. Notable RGB SOD approaches include deeply supervised networks with short connections [20], pyramid attention mechanisms for multi-scale feature enhancement [21], and edge-guided detection strategies [22]. Recent methods [23, 24] have achieved substantial progress by fusing multi-scale contextual features through various refinement strategies. However, RGB-based methods struggle under challenging conditions, including low illumination, intense noise, and complex backgrounds, necessitating exploration of auxiliary modalities.

To address RGB limitations, depth information was introduced as a complementary modality to provide 3D structural and spatial layout information [6]. Mainstream RGB-D methods adopt cross-modal fusion strategies to integrate RGB and depth features at multiple scales. Representative approaches include depth-aware multi-scale weighting [25], adaptive feature fusion strategies [26], joint learning with dense collaboration [27], and cross-modal attention mechanisms [28]. Despite these advances, depth sensors suffer from poor imaging quality under adverse conditions such as insufficient illumination and bad weather, limiting their practical deployment in real-world scenarios.

2.2 RGB-T SOD

RGB-Thermal SOD leverages thermal infrared sensors that capture radiation emitted by objects, offering robustness to lighting variations and effectively highlighting object contours in challenging environments. Early RGB-T methods relied on graph-based techniques with hand-crafted features. Wang et al. [10] established the first RGB-T benchmark (VT821) using multi-task manifold ranking, while Tu et al. [31] introduced collaborative graph learning and created the VT1000 dataset. With the advent of deep learning, CNN-based methods achieved significant advances through specialized cross-modal fusion mechanisms. Representative works include context-guided fusion modules [29], cross-guided fusion networks with self-attention [30], multi-interactive dual-stream decoders [31], and transformer architectures [32]. Recent approaches have explored modality difference mitigation [33], weighted fusion schemes [15], and prototype-based cross-modal integration [12].

Despite these advances, existing methods face critical limitations. Most approaches follow a Cross-Modal then Cross-Scale paradigm, directly fusing multi-modal features without adequate intra-modal semantic enhancement, resulting in suboptimal saliency instance representation. Additionally, the semantic gap between thermal modality and saliency attributes remains underexplored. Unlike depth maps, thermal images reflect temperature distributions without an inherent correlation to saliency, yet current methods fail to clearly define this relationship or fundamentally improve the original feature quality. Unlike existing approaches, this work enhances intra-modal representations by hierarchical multi-scale context aggregation and global dependency modeling before fusion. We introduce a cross-interactive dual attention mechanism that processes complementary feature streams through spatial and channel pathways with bidirectional information exchange, enabling comprehensive feature refinement while maintaining robustness to modality deficiencies and domain discrepancies.

3 Proposed Methodology

3.1 Overview

The proposed RGB-T SOD framework can be understood as a two-sensor cooperation pipeline. It first learns strong representations from RGB and thermal images separately, then combines them at multiple scales, and finally applies attention to emphasize the most informative regions and feature channels, which improves robustness when one modality is less reliable.

The proposed RGB-T SOD framework employs a dual-stream encoder–decoder architecture that

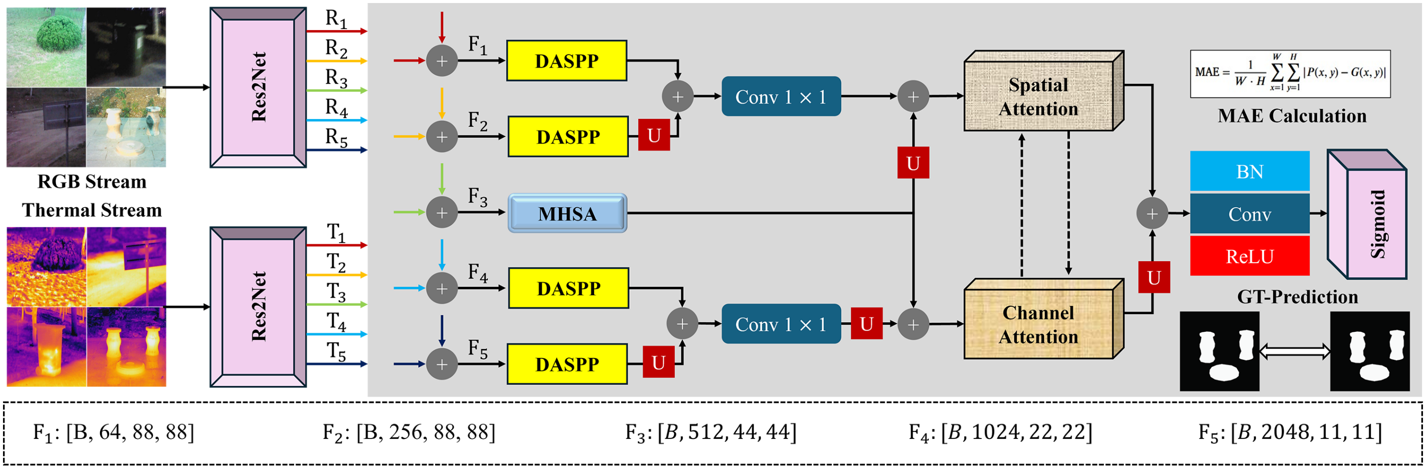


Figure 1. Overall architecture of PACNet. The framework employs dual backbones to extract hierarchical features from RGB and thermal modalities, followed by DASPP, MHSA, and a cross-interactive dual attention mechanism for feature refinement. Feature dimensions are shown at the bottom.

leverages complementary information from RGB and thermal modalities through hierarchical feature integration and a dual-attention mechanism. As illustrated in Figure 1, the network processes RGB and thermal images through parallel Res2Net backbone encoders to extract multi-scale hierarchical features. These features are then processed through DASPP and MHSA modules to capture multi-scale and global context, followed by strategic feature fusion at multiple scales. The architecture culminates in a cross-interactive dual-attention mechanism that processes two distinct feature streams via spatial and channel attention pathways, enabling effective cross-modal collaboration for robust salient object detection.

3.2 Multi-Modal Feature Extraction and Fusion

The feature extraction stage employs two parallel Res2Net-50 backbones to process the RGB and thermal input streams independently. Given an input RGB image $I_{RGB} \in \mathbb{R}^{H\times W\times 3}$ and its corresponding thermal image $I_T \in \mathbb{R}^{H\times W\times 1}$, each Res2Net encoder extracts five hierarchical feature representations $\{R_i\}_{i=1}^{5}$ and $\{T_i\}_{i=1}^{5}$ respectively, where each level captures progressively higher-level semantic information with reduced spatial resolution. The Res2Net architecture enhances multi-scale representational capability through hierarchical residual-like connections within individual blocks, making it particularly effective for capturing diverse scale variations inherent in RGB-T saliency detection scenarios. The feature dimensions at each hierarchical level are: $F_1 \in \mathbb{R}^{B\times 64\times 88\times 88}$, $F_2 \in \mathbb{R}^{B\times 256\times 88\times 88}$, $F_3 \in \mathbb{R}^{B\times 512\times 44\times 44}$, $F_4 \in \mathbb{R}^{B\times 1024\times 22\times 22}$, and $F_5 \in \mathbb{R}^{B\times 2048\times 11\times 11}$, where B denotes the batch size. The hierarchical features from both modalities are fused through element-wise addition at each level, producing integrated multi-modal representations:

$$F_i = R_i + T_i, \quad \text{for } i \in \{1, 2, 3, 4, 5\} \tag{1}$$

This early fusion strategy enables the network to combine complementary information from both modalities while maintaining computational efficiency and preserving the distinct characteristics of each semantic level. The additive fusion allows thermal features to supplement RGB features by providing additional cues in challenging scenarios such as low-light conditions or camouflaged objects.

3.3 DASPP Module

To capture multi-scale contextual information and expand the receptive field without sacrificing spatial resolution. DASPP modules are applied to four hierarchical levels: F_1, F_2, F_4, and F_5. The DASPP module applies parallel atrous convolutions with multiple dilation rates to capture objects and contexts at different scales simultaneously. For a given fused feature map F_i, the DASPP module generates scale-enriched representations through:

$$F^{i}_{DASPP} = \text{Concat}\left[\text{Conv}_{d_1}(F_i), \text{Conv}_{d_2}(F_i), \ldots, \text{Conv}_{d_n}(F_i)\right] \tag{2}$$

where Conv_{d_j} denotes atrous convolution with dilation rate d_j, and the outputs are concatenated along the channel dimension. The varying dilation rates enable the network to capture both fine-grained local details and broader contextual information, which are crucial for detecting salient objects at different scales in RGB and thermal imagery. The DASPP-processed features are denoted as F^1_{DASPP}, F^2_{DASPP}, F^4_{DASPP}, and F^5_{DASPP} for subsequent processing.

3.4 MHSA for Global Context Modeling

At the middle hierarchical level (F_3), we employ a MHSA mechanism to model long-range dependencies and capture global contextual relationships within the fused features. Unlike the DASPP modules that focus on multi-scale local-to-regional contexts. The MHSA mechanism also enable the network to establish relationships between spatially distant regions, which is essential for detecting salient objects with complex structures, irregular shapes, and multiple disconnected components. The MHSA module processes the fused feature F_3 by projecting it into query (Q), key (K), and value (V) representations through learned linear transformations:

$$Q = F_3W_Q, \quad K = F_3W_K, \quad V = F_3W_V \tag{3}$$

where W_Q, W_K, and W_V are learnable projection matrices. The self-attention mechanism computes the weighted feature representation as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

where d_k represents the dimension of the key vectors, and the scaling factor $\sqrt{d_k}$ prevents the dot products from becoming excessively large. By employing multiple attention heads, the MHSA module captures diverse relational patterns and semantic dependencies across different representation subspaces. In result, a robust cross-modal feature representations with enhanced global context awareness is obtained. The output is denoted F^3_{MHSA} and serves as a global context anchor for subsequent multi-scale feature integration.

3.5 Hierarchical Feature Integration

Following the DASPP and MHSA processing stages, the framework performs hierarchical feature integration through two parallel pathways that combine features from different semantic levels. This strategic integration enables the network to leverage both fine-grained spatial details and high-level semantic information.

Low to Middle Level Integration: The DASPP processed features from the first two levels are fused through upsampling and element-wise addition to create a unified low-level representation:

$$F_{12} = F^1_{DASPP} + \text{Upsample}(F^2_{DASPP}) \tag{5}$$

where the upsampling operation employs bilinear interpolation to match spatial resolutions. This fused feature F_{12} is then passed through a 1×1 convolution for channel reduction and feature transformation:

$$F'_{12} = \text{Conv}_{1\times1}(F_{12}) \tag{6}$$

The transformed low-level features are subsequently integrated with the global context from the MHSA module:

$$F_{123} = F'_{12} + \text{Upsample}(F^3_{MHSA}) \tag{7}$$

where F^3_{MHSA} is upsampled to match the spatial dimensions of F'_{12}. This integration produces a feature representation that combines fine-grained spatial details with global contextual information.

High Level Integration: Similarly, the DASPP processed features from the deeper levels are integrated to capture high-level semantic information:

$$F_{45} = \text{Upsample}(F^5_{DASPP}) + F^4_{DASPP} \tag{8}$$

The fused high-level features undergo channel reduction through 1×1 convolution:

$$F'_{45} = \text{Conv}_{1\times1}(F_{45}) \tag{9}$$

These transformed features are then enriched with global context:

$$F_{345} = F'_{45} + F^3_{MHSA} \tag{10}$$

where F^3_{MHSA} is spatially aligned with F'_{45} through appropriate upsampling or downsampling operations. The resulting feature F_{345} encapsulates high-level semantic information augmented with global contextual understanding.

3.6 Cross-Interactive Dual Attention Mechanism

The hierarchically integrated features F_{123} and F_{345} are processed through a cross-interactive dual attention mechanism that simultaneously performs spatial and channel-wise feature refinement. This design enables complementary feature enhancement through two parallel attention pathways with bidirectional information exchange.

Spatial Attention Pathway: The feature F_{123}, which encodes fine-grained spatial details with global context, is processed through the spatial attention module (Figure 2). The spatial attention mechanism computes attention weights across spatial locations

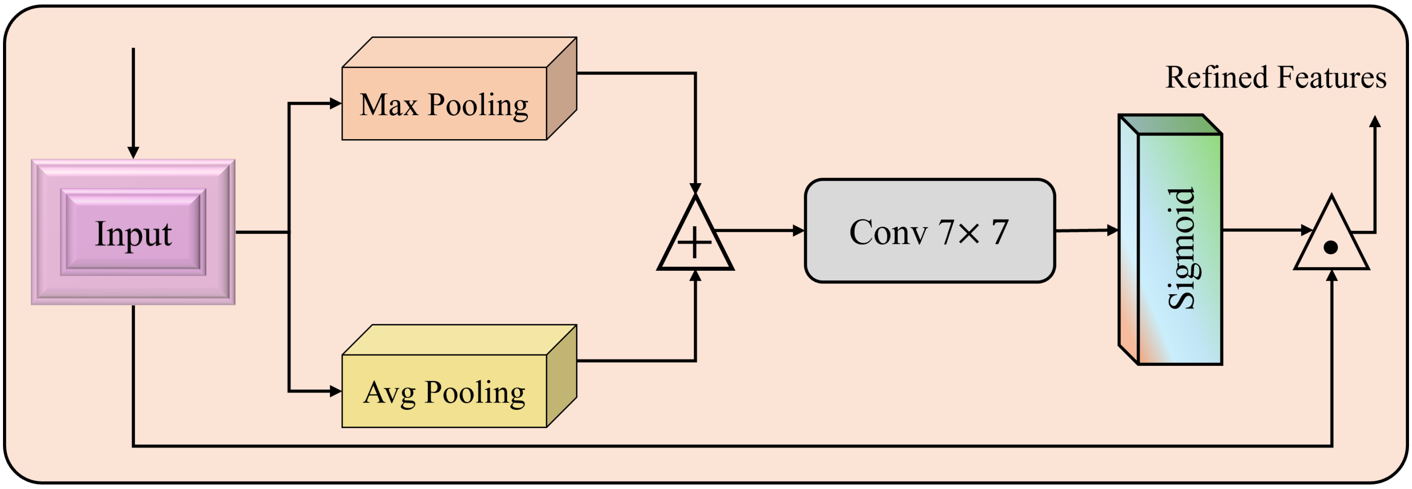


Figure 2. Feature flow inside spatial attention mechanism.

by aggregating channel information through parallel pooling operations:

$$A_{spatial}(F_{123}) = \sigma\Big(\text{Conv}_{7\times 7}([\text{AvgPool}(F_{123}); \text{MaxPool}(F_{123})])\Big) \tag{11}$$

$$F_{spatial} = F_{123} \otimes A_{spatial}(F_{123}) \tag{12}$$

where $\otimes$ denotes element-wise multiplication, σ represents the sigmoid activation functions. The average-pooled and max-pooled features are concatenated along the channel dimension before processing through a 7×7 convolution. This mechanism emphasizes spatially informative regions while suppressing background clutter.

Channel Attention Pathway: Concurrently, the feature F_{345}, which captures high-level semantic information, is processed through the channel attention module. The channel attention mechanism recalibrate channel-wise feature responses through global context aggregation:

$$A_{channel}(F_{345}) = \sigma\left(\text{MLP}\left(\text{GAP}(F_{345})\right)\right) \tag{13}$$

$$F_{channel} = F_{345} \otimes A_{channel}(F_{345}) \tag{14}$$

where GAP denotes global average pooling, and MLP represents a two-layer perceptron with reduction ratio r and ReLU activation that learns non-linear channel-wise dependencies. The channel attention adaptively emphasizes discriminative channels contributing to saliency detection.

Cross-Interaction: To enable mutual reinforcement between spatial and channel attention pathways, the architecture incorporates cross-interactive connections that allow each attention mechanism to influence the other:

$$F_{s\to c} = A_{channel}(F_{spatial}) \otimes F_{spatial} \tag{15}$$

$$F_{c\to s} = A_{spatial}(F_{channel}) \otimes F_{channel} \tag{16}$$

where $F_{s\to c}$ represents spatially-attended features refined by channel attention, and $F_{c\to s}$ represents channel-attended features refined by spatial attention. This bidirectional exchange ensures that both spatial localization and semantic channel information are optimally integrated.

3.7 Feature Concatenation and Saliency Prediction

The cross-interactive dual attention outputs are concatenated along the channel dimension to form a comprehensive feature representation:

$$F_{refined} = \text{Concat}\left[F_{s\to c}, F_{c\to s}\right] \tag{17}$$

The concatenated features are processed through a prediction head consisting of batch normalization (BN), convolution, and ReLU activation to generate the final saliency representation:

$$F_{pred} = \text{ReLU}\left(\text{Conv}\left(\text{BN}(F_{refined})\right)\right) \tag{18}$$

A sigmoid activation function is applied to produce the final saliency map $P(x, y) \in [0, 1]$, representing the pixel-wise probability of belonging to a salient object:

$$P(x, y) = \sigma(F_{pred}) \tag{19}$$

3.8 Loss Function

The network is trained using binary cross entropy (BCE) loss as the optimization objective, which is well-suited for pixel-wise binary classification tasks in saliency detection. The BCE loss measures the cross-entropy between the predicted saliency probabilities and the ground truth binary masks:

$$\mathcal{L}_{BCE} = - \frac{1}{W \cdot H} \sum_{x=1}^{W} \sum_{y=1}^{H} [G(x, y) \log(P(x, y)) + (1 - G(x, y)) \log(1 - P(x, y))] \tag{20}$$

where $P(x, y) \in [0, 1]$ denotes the predicted saliency probability at position (x, y), $G(x, y) \in \{0, 1\}$ represents the ground truth binary mask, and W and H are the width and height of the saliency map respectively. The BCE loss encourages the network to produce confident predictions by penalizing deviations from the ground truth labels at each pixel location. This formulation is particularly effective for RGB-T saliency detection as it provides strong gradient signals for both salient object pixels ($G(x, y) = 1$) and background pixels ($G(x, y) = 0$), enabling the network to learn clear decision boundaries between foreground and background regions across both modalities while maintaining robustness to the inherent class imbalance present in RGB-T saliency datasets.

4 Results and Discussion

4.1 Experimental Setup

Datasets We conduct extensive experiments across three widely used RGB-T benchmarks to comprehensively evaluate the proposed framework's performance under various challenging scenarios. The datasets include: (1) **VT821** [10], comprising 821 manually registered RGB-thermal image pairs with diverse challenging scenarios including occlusion, low illumination, and thermal crossover; (2) **VT1000** [11], consisting of 1,000 pairs captured by well-aligned RGB and thermal cameras with varied lighting conditions and scene complexities; and (3) **VT5000** [34], offering 5,000 pairs of high-resolution images with rich scene diversity, multiple object scales, and minimal spatial misalignment. Following the standard training protocol established in prior works [31], we utilize 2,500 image pairs from VT5000 for training, while the remaining 2,500 pairs from VT5000, along with all images from VT821 and VT1000, are reserved for testing. This data split ensures comprehensive evaluation across different dataset characteristics and acquisition conditions.

Evaluation Metrics We adopt five widely-used evaluation metrics for comprehensive performance assessment: (1) **S-measure** (S_m) evaluates structural similarity between predictions and ground truth, capturing region-aware and object-aware structural information; (2) **F-measure** (F_β) computes the weighted harmonic mean of precision and recall with $\beta^2 = 0.3$ to emphasize precision; (3) **weighted F-measure** (F_w^ξ) addresses the interpolation flaw in F_β by weighting errors based on their positions; (4) **E-measure** (E_m) jointly captures local pixel values and image-level mean to evaluate both pixel-level matching and global statistics; and (5) **Mean Absolute Error** ($\mathcal{M}$) measures average pixel-wise absolute difference between predictions and ground truth. For F_β and E_m, we report maximum values across all thresholds, while for F_w^ξ we report adaptive threshold values. Higher values of S_m, F_β, F_w^ξ, E_m, and lower values of $\mathcal{M}$ indicate superior performance.

Implementation Details The proposed framework is implemented in PyTorch 2.0 and trained on a single NVIDIA RTX 3090 GPU (24GB). The dual-stream Res2Net-50 backbone encoders for RGB and thermal modalities are initialized with ImageNet-pretrained weights, while newly introduced modules (DASPP, MHSA, dual-attention blocks, and prediction heads) are initialized using PyTorch's default Kaiming initialization. Input RGB and thermal images are resized to 352×352, with aspect ratio preserved via padding when necessary. The network is optimized end-to-end using AdamW $\beta_1 = 0.9$ and $\beta_2 = 0.999$ with weight decay of 10^{-4} and learning rate of 10^{-4}. The batch size is 8, and the model is trained for 150 epochs using binary cross-entropy (BCE) loss.

Key Hyperparameters for Reproducibility In the DASPP module, the main atrous dilation rates are set to $1, 3, 5, 7$. For MHSA, we use 8 attention heads with a head dimension consistent with the backbone feature width. In the channel-attention branch, the reduction ratio is set to r = 16, and the same ratio is used wherever squeeze-and-excitation style channel compression is applied.

Backbone Choice and Sensitivity we adopt Res2Net-50 for both modalities due to its strong multi-scale feature extraction ability, which benefits RGB-T saliency detection. The proposed PACNet modules are backbone-agnostic and can be applied to other encoders (e.g., ResNet-50 or lightweight backbones) by replacing the feature extractor while keeping the remaining architecture unchanged.

4.2 Comparison with State-of-the-Art Methods

We conduct comprehensive comparisons with SOTA RGB-T methods spanning from 2018 to 2023, including

Table 1. Quantitative comparison with state-of-the-art RGB-T SOD methods on VT821, VT1000, and VT5000 datasets. Best results are highlighted in **bold**. ↑ indicates higher is better, ↓ indicates lower is better.

Method	Year	VT821					VT1000					VT5000				
		$S_m\uparrow$	$F_\beta\uparrow$	$F_w^\xi\uparrow$	$E_m\uparrow$	$\mathcal{M}\downarrow$	$S_m\uparrow$	$F_\beta\uparrow$	$F_w^\xi\uparrow$	$E_m\uparrow$	$\mathcal{M}\downarrow$	$S_m\uparrow$	$F_\beta\uparrow$	$F_w^\xi\uparrow$	$E_m\uparrow$	$\mathcal{M}\downarrow$
MTMR	2018	0.725	0.662	0.462	0.815	0.108	0.706	0.715	0.485	0.836	0.119	0.680	0.595	0.387	0.795	0.114
SGDL	2020	0.765	0.730	0.583	0.847	0.085	0.787	0.764	0.652	0.856	0.090	0.750	0.672	0.558	0.824	0.089
ADF	2020	0.810	0.716	0.626	0.842	0.077	0.910	0.847	0.804	0.921	0.034	0.863	0.778	0.722	0.891	0.046
MIDD	2021	0.871	0.804	0.760	0.895	0.045	0.907	0.871	0.848	0.928	0.029	0.856	0.789	0.753	0.891	0.046
CSRNet	2021	0.885	0.830	**0.821**	0.908	0.038	0.918	0.877	0.878	0.925	0.024	0.868	0.810	0.796	0.905	0.042
MIA	2022	0.844	0.740	0.720	0.850	0.070	0.924	0.868	0.864	0.926	0.025	0.878	0.793	0.780	0.893	0.040
ECFFNet	2022	0.877	0.810	0.801	0.902	0.034	0.923	0.876	0.885	0.930	**0.021**	0.874	0.806	0.801	0.906	0.038
OSRNet	2022	0.875	0.813	0.801	0.896	0.043	**0.926**	0.892	0.891	0.935	0.022	0.875	0.823	0.807	0.908	0.040
LSNet	2023	0.878	0.825	0.809	0.911	0.033	0.925	0.885	0.887	0.935	0.023	0.877	0.825	0.806	0.915	0.037
Ours	2025	**0.886**	**0.832**	**0.821**	**0.912**	**0.032**	**0.926**	**0.893**	**0.894**	**0.937**	**0.021**	**0.880**	**0.828**	**0.807**	**0.917**	**0.034**

MTMR [10], SGDL [11], ADF [34], MIDD [31], CSRNet [29], MIA [35], ECFFNet [13], MMNet [36], OSRNet [37], and LSNet [38]. These methods represent diverse architectural paradigms, including early fusion approaches, attention-based mechanisms, and transformer-based architectures, providing a comprehensive benchmark for evaluating the proposed framework.

4.2.1 *Quantitative Analysis*

Table 1 presents quantitative comparisons across three benchmark datasets: VT821, VT1000, and VT5000. The proposed method achieves superior or competitive performance across all evaluation metrics and datasets, demonstrating its robustness and effectiveness in handling diverse challenging scenarios.

Performance on VT821: On the VT821 dataset, which contains diverse challenging scenarios including occlusion, low illumination, and thermal crossover, our method achieves the best performance across all metrics with $S_m = 0.886$, $F_\beta = 0.832$, $F_w^\xi = 0.821$, $E_m = 0.912$, and $\mathcal{M} = 0.032$. Compared to the second-best performing method CSRNet, our approach demonstrates improvements in S_m and F_β, while achieving comparable performance on F_w^ξ. Notably, our method outperforms LSNet on S_m, demonstrating the effectiveness of our hierarchical feature integration strategy combined with DASPP modules for multi-scale context aggregation.

Performance on VT1000: On the VT1000 dataset, which features well-aligned RGB-thermal pairs with varied lighting conditions, our method achieves superior results with $S_m = 0.926$ (tied with OSRNet), $F_\beta = 0.893$, $F_w^\xi = 0.894$, and $E_m = 0.937$, while matching the best MAE score of 0.021 (tied with ECFFNet). The exceptional performance on E_m (0.937) demonstrates our method's ability to capture both local pixel-level details and global image statistics effectively, which is crucial for handling the diverse object scales and scene complexities present in VT1000.

Performance on VT5000: On the most significant benchmark VT5000, which contains high-resolution image pairs with rich scene diversity, our method achieves the best performance across all metrics: $S_m = 0.880$, $F_\beta = 0.828$, $F_w^\xi = 0.807$, $E_m = 0.917$, and $\mathcal{M} = 0.034$. These consistent improvements demonstrate the robustness and scalability of our approach when handling large-scale datasets with diverse imaging conditions. The superior performance on VT5000 validates the effectiveness of our hierarchical feature integration strategy, where features F_{123} and F_{345} are constructed through strategic fusion with global context from MHSA, enabling the network to handle multiple object scales and complex backgrounds effectively.

4.2.2 *Qualitative Analysis*

Figure 3 presents visual comparisons between our method and representative SOTA approaches. The qualitative results demonstrate our method's superior ability to handle a range of challenging scenarios commonly encountered in RGB-T SOD. In low-contrast scenarios with subtle boundaries (Row 1), our method produces complete and accurate predictions, whereas competing methods generate incomplete detections with missing regions or inaccurate boundaries. When dealing with multiple objects at varying scales (Row 2), our hierarchical feature integration strategy effectively detects all objects with clear separation. For complex backgrounds with thermal clutter (Row 3), our cross-interactive dual attention

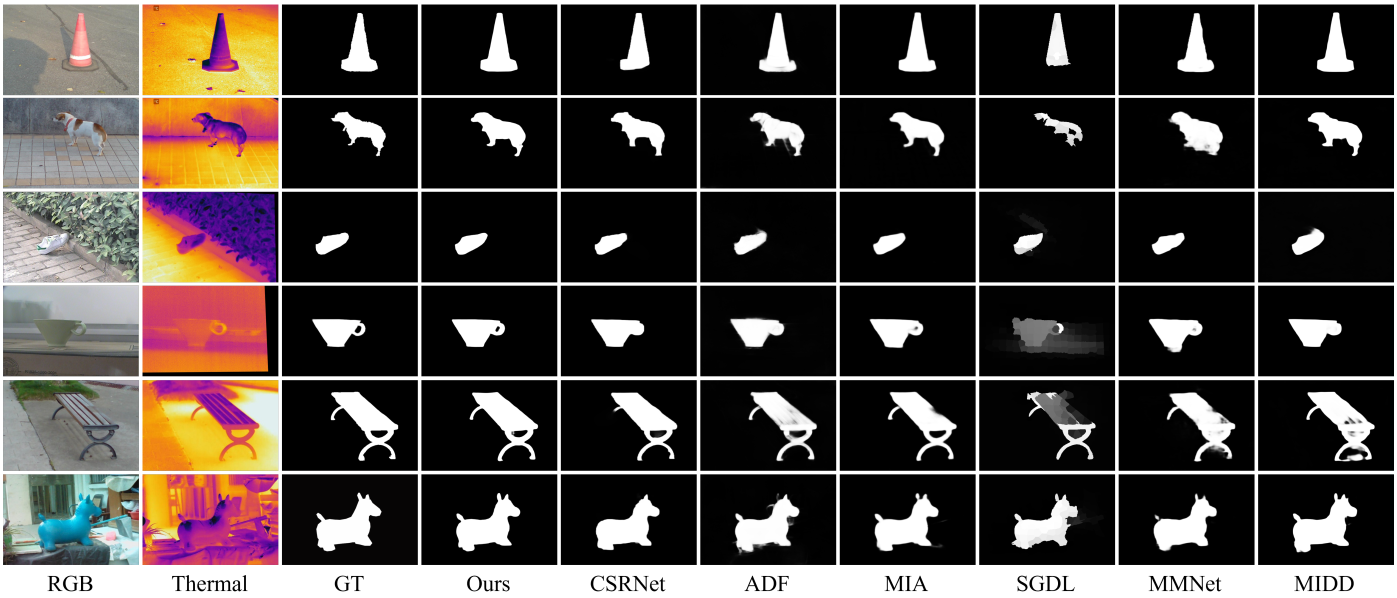


Figure 3. Qualitative comparison of state-of-the-art RGB–T SOD methods under representative challenging scenarios. From left to right: RGB image, thermal image, ground truth (GT), and predictions. Each row illustrates a specific challenge: Row 1—small object with weak thermal contrast; Row 2—fast-moving object with motion blur; Row 3—camouflaged object in a complex background; Row 4—low-contrast object in cluttered scenes; Row 5—thin-structured object with shape deformation; Row 6—multiple objects with varying scales and thermal ambiguity.

Table 2. Component-wise ablation study on VT821 and VT1000 datasets. Each row progressively adds components to validate their individual contributions.

Configuration	VT821					VT1000				
	$S_m \uparrow$	$F_\beta \uparrow$	$F_w^\xi \uparrow$	$E_m \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	$F_w^\xi \uparrow$	$E_m \uparrow$	$\mathcal{M} \downarrow$
Baseline	0.848	0.782	0.771	0.881	0.051	0.899	0.843	0.848	0.912	0.037
+ DASPP	0.862	0.801	0.788	0.893	0.044	0.909	0.861	0.864	0.921	0.031
+ MHSA	0.871	0.813	0.797	0.900	0.039	0.916	0.872	0.875	0.927	0.027
+ Spatial Attn	0.879	0.823	0.809	0.906	0.035	0.921	0.884	0.886	0.932	0.024
+ Channel Attn	0.883	0.828	0.816	0.909	0.033	0.924	0.889	0.890	0.935	0.022
+ Cross-interaction	**0.886**	**0.832**	**0.821**	**0.912**	**0.032**	**0.926**	**0.893**	**0.894**	**0.937**	**0.021**

mechanism achieves cleaner predictions with minimal false positives by effectively distinguishing genuine salient objects from thermal noise. At the same time, other methods struggle with scattered false positives and over-segmentation. Our method successfully preserves fine structural details, such as thin appendages and intricate boundaries (Row 4), by applying spatial attention to low-level features F_{123}, outperforming methods that produce coarse predictions with lost details. Under extremely low illumination conditions (Row 5), our method demonstrates robust performance by effectively leveraging thermal information through channel attention while compensating for unreliable RGB features, generating accurate predictions comparable to ground truth, where competing methods fail significantly.

4.3 Ablation Study

To validate the effectiveness of each proposed component, we conduct comprehensive ablation studies on the VT821 and VT1000 datasets. We systematically analyze the contributions of DASPP modules, MHSA mechanism, dual-attention components, and cross-interaction strategy.

4.3.1 Component-wise Ablation Analysis

Table 2 presents a systematic component-wise ablation study that progressively adds each proposed module to evaluate its individual contribution. The baseline model employs simple element-wise addition for multi-modal fusion with standard

Table 3. Ablation study on DASPP module placement across hierarchical levels on VT821 dataset.

DASPP Placement	$S_m \uparrow$	$F_\beta \uparrow$	$F_w^\xi \uparrow$	$E_m \uparrow$	$\mathcal{M} \downarrow$
Low-level only (F_1, F_2)	0.867	0.809	0.795	0.897	0.041
High-level only (F_4, F_5)	0.873	0.817	0.803	0.902	0.037
All levels $(F_1, F_2, F_3, F_4, F_5)$	0.880	0.825	0.813	0.907	0.034
Selective + MHSA at F_3 (Ours)	**0.886**	**0.832**	**0.821**	**0.912**	**0.032**

Table 4. Ablation study comparing different dual attention configurations on VT821 dataset.

Attention Configuration	$S_m \uparrow$	$F_\beta \uparrow$	$F_w^\xi \uparrow$	$E_m \uparrow$	$\mathcal{M} \downarrow$
No attention	0.871	0.813	0.797	0.900	0.039
Spatial attention only	0.879	0.823	0.809	0.906	0.035
Channel attention only	0.877	0.820	0.805	0.904	0.036
Both without cross-interaction	0.883	0.828	0.816	0.909	0.033
Both with cross-interaction (Ours)	**0.886**	**0.832**	**0.821**	**0.912**	**0.032**

convolutions, achieving $S_m = 0.848$ and $F_\beta = 0.782$ on VT821. Adding DASPP modules to levels F_1, F_2, F_4, F_5 substantially improves performance to $S_m = 0.862$ and $F_\beta = 0.801$, demonstrating the importance of multi-scale context aggregation through atrous convolutions with varying dilation rates. Incorporating the MHSA module at F_3 for global context modeling further enhances performance to $S_m = 0.871$ and $F_\beta = 0.813$, validating the benefit of capturing long-range dependencies and global contextual relationships.

Adding spatial attention to process F_{123} features yields significant improvements ($S_m = 0.879$, $F_\beta = 0.823$), indicating the effectiveness of emphasizing spatially informative regions for fine-grained localization. Incorporating channel attention for F_{345} features provides an additional performance boost to $S_m = 0.883$ and $F_\beta = 0.828$, demonstrating the value of channel-wise semantic recalibration for high-level feature refinement. Finally, introducing cross-interaction between spatial and channel attention pathways achieves the best performance with $S_m = 0.886$ and $F_\beta = 0.832$, validating our hypothesis that bidirectional information exchange between attention mechanisms enables more comprehensive feature refinement.

4.3.2 *DASPP Module Placement Analysis*

Table 3 examines optimal DASPP placement across hierarchy levels. We test four configurations: DASPP on low-level (F_1, F_2), high-level (F_4, F_5), all levels $(F_1\text{-}F_5)$, and our selective approach $(F_1, F_2, F_4, F_5$ with MHSA at $F_3)$. DASPP on low-level features achieves $S_m = 0.867$, capturing details but lacking high-level context. High-level DASPP improves to $S_m = 0.873$, benefiting from semantic multi-scale aggregation but missing spatial details. All-level DASPP yields $S_m = 0.880$, showing consistent multi-scale processing. Our selective method, combining DASPP at F_1, F_2, F_4, F_5 with MHSA at F_3, outperforms others with $S_m = 0.886$ and $F_\beta = 0.832$. It balances local-to-regional multi-scale contexts and models global dependencies, demonstrating that combining DASPP and MHSA is more effective than applying a single mechanism uniformly across all levels.

4.3.3 *Dual Attention Mechanism Analysis*

Table 4 analyzes different configurations of the dual attention mechanism. The baseline without attention mechanisms achieves $S_m = 0.871$ and $F_\beta = 0.813$. Applying spatial attention alone to F_{123} improves performance to $S_m = 0.879$ and $F_\beta = 0.823$, demonstrating the effectiveness of emphasizing spatially informative regions for accurate localization. Using only channel attention on F_{345} yields $S_m = 0.877$ and $F_\beta = 0.820$, showing the benefit of channel-wise semantic recalibration, though slightly lower than spatial attention alone. Employing both attention mechanisms in parallel without cross-interaction achieves $S_m = 0.883$ and $F_\beta = 0.828$, indicating that spatial and channel attention provide complementary information when applied to different feature streams. Finally, our proposed cross-interactive dual attention achieves the best performance with $S_m = 0.886$ and $F_\beta = 0.832$. The cross-interaction mechanism enables bidirectional information exchange: spatial attention features are refined by channel attention and vice versa,

Table 5. Ablation study on different hierarchical feature integration strategies on VT821 dataset.

Integration Strategy	$S_m \uparrow$	$F_\beta \uparrow$	$F_w^\xi \uparrow$	$E_m \uparrow$	$\mathcal{M} \downarrow$
Direct concatenation	0.859	0.796	0.783	0.889	0.046
Progressive fusion (FPN-style)	0.873	0.815	0.801	0.901	0.038
Two-stream without F_3	0.878	0.821	0.808	0.905	0.035
Two-stream with F_3 (Ours)	**0.886**	**0.832**	**0.821**	**0.912**	**0.032**

resulting in more comprehensive feature refinement than independent parallel processing. understanding.

4.3.4 *Feature Integration Strategy Analysis*

Table 5 evaluates different strategies for integrating hierarchical features. Direct concatenation of all DASPP processed features achieves $S_m = 0.859$ and $F_\beta = 0.796$, demonstrating that simple concatenation without strategic integration is suboptimal. Progressive fusion, following a FPN style, improves performance to $S_m = 0.873$ and $F_\beta = 0.815$, demonstrating the benefit of hierarchical feature aggregation. Our two-stream integration without incorporating F_3 global context (i.e., F_{12} and F_{45} only) achieves $S_m = 0.878$ and $F_\beta = 0.821$, validating the effectiveness of creating separate pathways for low-level and high-level features. Finally, our complete two-stream integration strategy that fuses both streams with F_3 global context (producing F_{123} and F_{345}) achieves the best performance with $S_m = 0.886$ and $F_\beta = 0.832$. This demonstrates that integrating global context from MHSA-processed F_3 into both feature streams is crucial. The strategic integration of global context enriches both pathways, enabling more effective cross-interactive dual attention processing.

5 Conclusion

This paper presented PACNet, an RGB–thermal salient object detection framework designed to address key limitations of existing RGB-T methods in multi-scale context modeling, feature refinement, and modality fusion. PACNet integrates hierarchical feature representations and introduces a cross-interactive dual-attention mechanism that strengthens intra-modal features prior to fusion and enables bidirectional information exchange between spatial localization cues and channel-wise semantic responses. As a result, PACNet achieves state-of-the-art performance on three public benchmarks—VT821, VT1000, and VT5000—with S-measure scores of 0.886, 0.926, and 0.880, respectively. Ablation and comparative experiments further confirm that each module contributes meaningfully to performance gains, particularly in challenging conditions such as low illumination, thermal clutter, and multi-scale targets.

Beyond benchmark improvements, the proposed design offers a general and practical fusion paradigm for robust multi-modal perception, where complementary sensing (RGB and thermal) is required to handle degraded visual conditions. Moreover, PACNet's refinement-before-fusion design and attention-guided recalibration are intended to mitigate performance degradation when one modality is missing or corrupted, which motivates our planned robustness evaluation under controlled modality degradation settings. Future work will focus on developing lightweight and efficient variants of PACNet for real-time deployment, evaluating robustness under missing or corrupted modalities, and extending the framework to additional modalities (e.g., depth or event data) to further improve generalization in extreme environments. Future work will also evaluate cross-dataset generalization (train on one VT dataset and test on another) to quantify robustness under domain shift.

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] Wang, Y., Li, G., & Liu, Z. (2023). SGFNet: Semantic-guided fusion network for RGB-thermal semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology, 33*(12), 7737-7748. [CrossRef]

[2] Xu, X., Zhao, J., Wu, J., & Shen, F. (2022). Switch and refine: A long-term tracking and segmentation framework. *IEEE Transactions on Circuits and Systems for Video Technology, 33*(3), 1291-1304. [CrossRef]

[3] Yang, J., Wei, P., & Zheng, N. (2023). Cross time-frequency transformer for temporal action localization. *IEEE Transactions on Circuits and Systems for Video Technology, 34*(6), 4625-4638. [CrossRef]

[4] 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]

[5] Han, J., Zhang, D., Hu, X., Guo, L., Ren, J., & Wu, F. (2014). Background prior-based salient object detection via deep reconstruction residual. *IEEE Transactions on Circuits and Systems for Video Technology, 25*(8), 1309-1321. [CrossRef]

[6] 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]

[7] Hu, X., Sun, F., Sun, J., Wang, F., & Li, H. (2024). Cross-modal fusion and progressive decoding network for RGB-D salient object detection. *International Journal of Computer Vision, 132*(8), 3067-3085. [CrossRef]

[8] Chen, G., Shao, F., Chai, X., Chen, H., Jiang, Q., Meng, X., & Ho, Y.-S. (2022). CGMDRNet: Cross-guided modality difference reduction network for RGB-T salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology, 32*(9), 6308-6323. [CrossRef]

[9] Ma, Y., Sun, D., Meng, Q., Ding, Z., & Li, C. (2017, December). Learning multiscale deep features and SVM regressors for adaptive RGB-T saliency detection. In *2017 10th International Symposium on Computational Intelligence and Design (ISCID)* (Vol. 1, pp. 389-392). IEEE. [CrossRef]

[10] Wang, G., Li, C., Ma, Y., Zheng, A., Tang, J., & Luo, B. (2018). RGB-T saliency detection benchmark: Dataset, baselines, analysis and a novel approach. In *Chinese Conference on Image and Graphics Technologies* (pp. 359-369). Springer. [CrossRef]

[11] Tu, Z., Xia, T., Li, C., Wang, X., Ma, Y., & Tang, J. (2019). RGB-T image saliency detection via collaborative graph learning. *IEEE Transactions on Multimedia, 22*(1), 160-173. [CrossRef]

[12] Zhang, Z., Wang, J., & Han, Y. (2023). Saliency prototype for RGB-D and RGB-T salient object detection. In *Proceedings of the 31st ACM International Conference on Multimedia* (pp. 3696-3705). [CrossRef]

[13] Zhou, W., Guo, Q., Lei, J., Yu, L., & Hwang, J.-N. (2021). ECFFNet: Effective and consistent feature fusion network for RGB-T salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology, 32*(3), 1224-1235. [CrossRef]

[14] Tu, Z., Li, Z., Li, C., Lang, Y., & Tang, J. (2021). Multi-interactive dual-decoder for RGB-thermal salient object detection. *IEEE Transactions on Image Processing, 30*, 5678-5691. [CrossRef]

[15] Wang, K., Tu, Z., Li, C., Zhang, C., & Luo, B. (2024). Learning adaptive fusion bank for multi-modal salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology, 34*(8), 7344-7358. [CrossRef]

[16] 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]

[17] 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]

[18] Jiang, Z., & Davis, L. S. (2013). Submodular salient region detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2043-2050). [CrossRef]

[19] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770-778).

[20] Hou, Q., Cheng, M.-M., Hu, X., Borji, A., Tu, Z., & Torr, P. H. (2017). Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3203-3212).

[21] Zhang, X., Wang, T., Qi, J., Lu, H., & Wang, G. (2018). Progressive attention guided recurrent network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 714-722). [CrossRef]

[22] Zhao, J. X., Liu, J. J., Fan, D. P., Cao, Y., Yang, J., & Cheng, M. M. (2019). EGNet: Edge guidance network for salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 8779-8788). [CrossRef]

[23] Wei, J., Wang, S., & Huang, Q. (2020). F^{3}Net: Fusion, feedback and focus for salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 07, pp. 12321-12328). [CrossRef]

[24] 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]

[25] Piao, Y., Ji, W., Li, J., Zhang, M., & Lu, H. (2019). Depth-induced multi-scale recurrent attention network for saliency detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 7254-7263). [CrossRef]

[26] Li, C., Cong, R., Kwong, S., Hou, J., Fu, H., Zhu, G., Zhang, D., & Huang, Q. (2020). ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection. *IEEE Transactions on Cybernetics, 51*(1), 88-100. [CrossRef]

[27] Fu, K., Fan, D. P., Ji, G. P., & Zhao, Q. (2020). JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3052-3062). [CrossRef]

[28] Hu, X., Sun, F., Sun, J., Wang, F., & Li, H. (2024). Cross-modal fusion and progressive decoding network for RGB-D salient object detection. *International Journal of Computer Vision, 132*(8), 3067-3085. [CrossRef]

[29] Huo, F., Zhu, X., Zhang, L., Liu, Q., & Shu, Y. (2021). Efficient context-guided stacked refinement network for RGB-T salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology, 32*(5), 3111-3124. [CrossRef]

[30] Wang, J., Song, K., Bao, Y., Huang, L., & Yan, Y. (2021). CGFNet: Cross-guided fusion network for RGB-T salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology, 32*(5), 2949-2961. [CrossRef]

[31] Tu, Z., Li, Z., Li, C., Lang, Y., & Tang, J. (2021). Multi-interactive dual-decoder for RGB-thermal salient object detection. *IEEE Transactions on Image Processing, 30*, 5678-5691. [CrossRef]

[32] Liu, Z., Tan, Y., He, Q., & Xiao, Y. (2021). SwinNet: Swin transformer drives edge-aware RGB-D and RGB-T salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology, 32*(7), 4486-4497. [CrossRef]

[33] Cong, R., Zhang, K., Zhang, C., Zheng, F., Zhao, Y., Huang, Q., & Kwong, S. (2022). Does thermal really always matter for RGB-T salient object detection? *IEEE Transactions on Multimedia, 25*, 6971-6982. [CrossRef]

[34] Tu, Z., Ma, Y., Li, Z., Li, C., Xu, J., & Liu, Y. (2022). RGBT salient object detection: A large-scale dataset and benchmark. *IEEE Transactions on Multimedia, 25*, 4163-4176. [CrossRef]

[35] Liang, Y., Qin, G., Sun, M., Qin, J., Yan, J., & Zhang, Z. (2022). Multi-modal interactive attention and dual progressive decoding network for RGB-D/T salient object detection. *Neurocomputing, 490*, 132-145. [CrossRef]

[36] Gao, W., Liao, G., Ma, S., Li, G., Liang, Y., & Lin, W. (2021). Unified information fusion network for multi-modal RGB-D and RGB-T salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology, 32*(4), 2091-2106. [CrossRef]

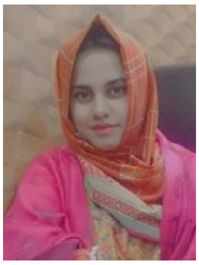
[37] Huo, F., Zhu, X., Zhang, Q., Liu, Z., & Yu, W. (2022). Real-time one-stream semantic-guided refinement network for RGB-thermal salient object detection. *IEEE Transactions on Instrumentation and Measurement, 71*, 1-12. [CrossRef]

[38] Zhou, W., Zhu, Y., Lei, J., Yang, R., & Yu, L. (2023). LSNet: Lightweight spatial boosting network for detecting salient objects in RGB-thermal images. *IEEE Transactions on Image Processing, 32*, 1329-1340. [CrossRef]

Muhammad Zain Hassan is an undergraduate student in the Software Engineering Department at the University of Haripur, Pakistan. His research interests include computer vision, image processing, deep learning, and multi-modal learning with a particular focus on salient object detection and semantic segmentation. He is actively involved in data annotation and collection for computer vision applications, contributing to dataset curation and benchmark development. His current work explores attention mechanisms and cross-modal fusion strategies for robust visual perception in challenging environments. (Email: zainhassanzero@gmail.com)

Alexandros Gazis received his diploma in Electronic and Computer Engineering and his MSc in Microelectronics and Computer Systems from the Department of Electrical and Computer Engineering, Democritus University of Thrace, Greece, in 2016 and 2018, respectively. Since 2018, he has been a PhD candidate in the field of computer science at the same university, where he is a member of the "Operating Systems and Middleware for Pervasive Computing and Wireless Sensor Networks" research group. He is also currently pursuing an MBA at Heriot-Watt University since February 2023. Moreover, he is a Teaching Assistant and Lab Demonstrator, supervised by Assistant Professor Eleftheria Katsiri. Mr. Gazis is a member of the Technical Chamber of Greece and works in the private sector as a Software Engineer for Piraeus Bank S.A., specializing in banking systems. He has published articles on Artificial Intelligence, game engines, web data analytics, remote sensing, and neural networks. His research focuses on the Internet of Things via wireless sensor networks, cloud computing, and middleware development for pervasive computing. (Email: agazis@ee.duth.gr)

Abdurrahman Khan Abdurrahman Khan is a Computer Science student currently expanding his skills in machine learning, Python programming, data structures, and software development, with a particular focus on strengthening his technical foundation. He is passionate about integrating technology and design to create innovative digital solutions that are both functional and visually coherent. His current focus is on building a strong foundation in programming and design tools, which he considers essential for preparing for a professional freelancing and research career in the technology field. (Email: iAbdurrahman3797@gmail.com)



Zainab Ghazanfar received her M.S. degree in Computer Science from the University of Lahore. She served as a Lecturer in the Department of Computer Science at the University of Management and Technology (UMT), Lahore. She has also held a lecturer position in the Department of Computer Science at Lahore Garrison University. Currently, she is pursuing a Ph.D. in the Department of Software and Artificial Intelligence at Gachon University, South Korea. Her research interests include deep learning, medical image processing, image recognition, and the application of artificial intelligence in medical images. (Email: zainab@gachon.ac.kr)