http://dx.doi.org/10.62762/TSCC.2025.390515

RESEARCH ARTICLE

MAFNet: Multi-level Attention Fusion Network for Precise Prominence Analysis in Visual Sensing Systems

Farhan Ali[1,*], **Muhammad Ali**[2] and **Zaid Muhammad**[3]

[1] Department of Computer Science, Graz University of Technology, Graz 8010, Austria
[2] Department of Software Engineering, University of Haripur, Haripur, Pakistan
[3] Global Degree College, Peshawar, Pakistan

Abstract

Salient object detection aims to identify and segment the most visually prominent objects in images. Despite significant advances in deep learning, existing methods struggle to balance global context modeling, boundary preservation, and multi-scale feature integration. To address these limitations, we propose MAFNet (Multi-level Attention Fusion Network), a novel attention-driven framework that leverages specialized attention mechanisms tailored to different semantic levels. Our approach employs a Tokens-to-Token (T2T) Transformer backbone for hierarchical feature extraction, capturing both local structural details and global contextual relationships. The core contribution lies in a comprehensive attention framework comprising six specialized modules: Contextual Feature Extraction (CFE) for multi-scale context refinement, Contour Aware Attention (CAA) for boundary preservation, Pyramidal Spatial Attention (PSA) for hierarchical spatial reasoning, Efficient Multi-Head Attention (EMHA) for semantic enhancement, Semantic Relation Attention (SRA) for global context modeling, and Frequency Channel Attention (FCA) for frequency-domain feature enhancement. These refined features are integrated through a parallel multi-path decoder that efficiently fuses information from different semantic levels. Extensive experiments on six benchmark datasets (ECSSD, PASCAL-S, SOD, DUTS-TE, HKU-IS, and DUT-OMRON) demonstrate that MAFNet achieves state-of-the-art performance, with particular strengths in handling complex object configurations and preserving fine-grained boundaries.

Keywords: saliency detection, multi-level attention, feature fusion, contour awareness, frequency channel attention.

1 Introduction

Humans possess a remarkable ability to quickly identify and focus on the most salient elements in complex scenes, a capacity rooted in biological mechanisms of visual attention. Salient Object Detection (SOD) seeks to replicate this cognitive process computationally by automatically identifying and segmenting the most eye-catching objects in images, a task that has been extensively benchmarked across diverse model families and challenging datasets [1]. This technology has become foundational

Submitted: 19 December 2025
Accepted: 07 February 2026
Published: 30 June 2026

Vol. 3, **No.** 2, 2026.
10.62762/TSCC.2025.390515

***Corresponding author:**
✉ Farhan Ali
farhan.ali@student.tugraz.at

Citation
Ali, F., Ali, M., & Muhammad, Z. (2026). MAFNet: Multi-level Attention Fusion Network for Precise Prominence Analysis in Visual Sensing Systems. *ICCK Transactions on Sensing, Communication, and Control*, 3(2), 124–138.

for various computer vision tasks, including object tracking, content-aware image manipulation, and intelligent segmentation systems.

SOD methodologies can be broadly divided into two categories: RGB-based and RGB-D-based techniques. Although RGB-D approaches leverage depth information for enhanced accuracy, their deployment in real-world scenarios is constrained by the reliance on high-quality depth sensors and the associated computational overhead for cross-modal fusion. In contrast, RGB-based methods[2] offer a more practical approach by extracting salient regions directly from color images. These approaches have made substantial progress through sophisticated feature representations and semantic understanding, effectively balancing the preservation of local detail with global context awareness.

Early SOD systems predominantly relied on handcrafted features and low-level visual cues [3, 4]. These traditional methods exploited characteristics such as edge information, textural patterns, and color distinctiveness, while incorporating design principles based on center-surround contrast and background modeling [5]. The computational framework employed bottom-up processing strategies that analyzed images through manually engineered features rather than data-driven learning. Despite their foundational importance, these approaches struggled to capture high-level semantic concepts, resulting in difficulties in accurately separating foreground objects from complex backgrounds, particularly under challenging conditions involving multiple objects or variable illumination.

The emergence of deep convolutional neural networks has revolutionized visual understanding tasks, introducing powerful mechanisms for automatic feature learning [6]. This progress has extended beyond single-modality RGB analysis: depth-aware fusion architectures have been developed to exploit complementary geometric cues for RGB-D salient object detection [7], while multi-modal fusion networks integrating visible, depth, and thermal cues have further demonstrated the value of cross-modal feature learning for robust saliency prediction under challenging imaging conditions [8]. This paradigm shift has proven equally transformative for SOD, with deep models demonstrating superior ability to capture intricate spatial dependencies [9–12]. Initial CNN-based SOD frameworks operated on region-based predictions, assigning uniform saliency scores across detected regions. Progressive refinements led to the adoption of fully convolutional architectures, which enable dense, pixel-wise saliency estimation through hierarchical feature extraction [13]. Subsequent innovations focused on sophisticated multi-scale feature integration strategies and on incorporating atrous convolutions to broaden receptive fields [14]. Contemporary research has emphasized attention mechanisms to enhance contextual fusion, with recent work proposing efficient attention-driven architectures that process multi-resolution features via optimized backbone networks [15, 16].

Despite significant progress in SOD research, existing approaches continue to face several fundamental challenges that limit their effectiveness. Traditional CNN-based methods struggle to capture comprehensive global contextual relationships because they inherently prioritize local information, which limits their ability to model scene-wide dependencies crucial for accurate salient object detection [16]. This limitation becomes particularly evident in complex scenarios in which understanding the relationships between distant regions is essential. Furthermore, most contemporary SOD models rely on generic attention mechanisms that fail to account for the distinct characteristics of features at different semantic levels. This uniform treatment of low-, medium-, and high-level representations leads to suboptimal feature refinement and computational inefficiency, because different feature scales inherently possess distinct spatial and semantic properties that require specialized processing. Additionally, current multi-scale feature integration strategies often employ direct fusion approaches without adequate processing mechanisms, resulting in feature degradation when combining information across different scales [10]. The challenge lies in developing an efficient progressive fusion strategy that preserves feature fidelity while effectively merging fine-grained details with high-level semantic information, particularly crucial for practical deployment scenarios where both accuracy and computational efficiency are paramount.

1.1 Contributions

To address the limitations above, we propose a novel attention-driven framework for efficient SOD that introduces several key contributions:

1. **Transformer-based Global Context Modeling:** We employ a Tokens-to-Token (T2T) Transformer as the backbone feature extractor to effectively

capture long-range dependencies and global contextual relationships across the entire image. This approach overcomes the local receptive-field constraints inherent in traditional CNN-based architectures, enabling comprehensive scene understanding and precise salient object localization.

2. **Specialized Multi-level Attention Mechanisms:** We introduce a comprehensive attention framework comprising six distinct modules tailored to different feature characteristics: Contextual Feature Extraction (CFE) modules for multi-scale feature refinement, Contour Aware Attention (CAA) for boundary preservation, Pyramidal Spatial Attention (PSA) for hierarchical spatial reasoning, Efficient Multi-Head Attention (EMHA) for semantic enhancement, Semantic Relation Attention (SRA) for global context modeling, and Frequency Channel Attention (FCA) for frequency-domain feature enhancement. This specialized design ensures optimal feature processing at different semantic levels.

3. **Parallel Multi-Path Feature Fusion Decoder:** We design an efficient decoder that fuses multi-scale features from three parallel attention-enhanced streams through direct integration. The decoder combines fine-grained spatial details from shallow layers with high-level semantic information from deeper layers via element-wise addition and upsampling. This parallel fusion strategy preserves complementary information across different semantic levels to enhance SOD performance.

4. **Comprehensive and Efficient Architecture:** Our complete framework achieves an optimal balance between detection accuracy and computational efficiency, making it suitable for practical deployment scenarios. The synergistic combination of transformer-based feature extraction, specialized attention mechanisms, and parallel feature integration delivers state-of-the-art (SOTA) performance on six SOD benchmark datasets.

2 Related Work

This section reviews the evolution of deep learning approaches to SOD, focusing on feature-extraction architectures, multi-scale integration strategies, and the integration of attention mechanisms with vision transformers.

2.1 Deep Neural Network Architectures for SOD

The advent of deep learning has fundamentally transformed SOD, enabling sophisticated feature extraction and representation learning. Early efforts focused on leveraging CNN architectures to capture multi-scale features and combine global and local contextual information to improve spatial coherence [17, 18]. Progressive developments emphasized pixel-level precision through specialized network designs. Hierarchical architectures were introduced to systematically refine saliency maps by integrating contextual cues at multiple levels [19], while recurrent mechanisms enabled iterative refinement through saliency priors [20]. Advanced approaches incorporated reverse attention mechanisms to address output resolution challenges and model complexity [13]. Bidirectional message-passing frameworks have emerged to capture diverse contextual information across feature map levels via multi-scale feature-extraction blocks [21]. Recent innovations have explored zero-shot learning paradigms that map salient and background features into a unified metric space, reducing dependence on large-scale training datasets while maintaining detection accuracy [22]. These developments reflect a continuous evolution toward architectures that effectively balance semantic understanding with computational efficiency.

2.2 Multi-Scale Feature Integration Strategies

Effective SOD requires sophisticated mechanisms to integrate features across multiple scales, combining high-level semantic information with fine-grained details [23]. Multi-level feature fusion techniques have been extensively explored through various architectural paradigms. Hierarchical frameworks employing stage-wise refinement have demonstrated effectiveness in capturing relative saliency [24]. Computationally efficient integration methods have been developed to maintain performance while reducing computational overhead [25]. Pyramid-based architectures have proven particularly effective, incorporating context-dependent feature-extraction modules and attention mechanisms that operate on both the channel and spatial dimensions [14]. These approaches enhance feature quality across levels while preserving boundary information through specialized loss functions. Dense connectivity patterns have been leveraged to improve feature integration and attention map

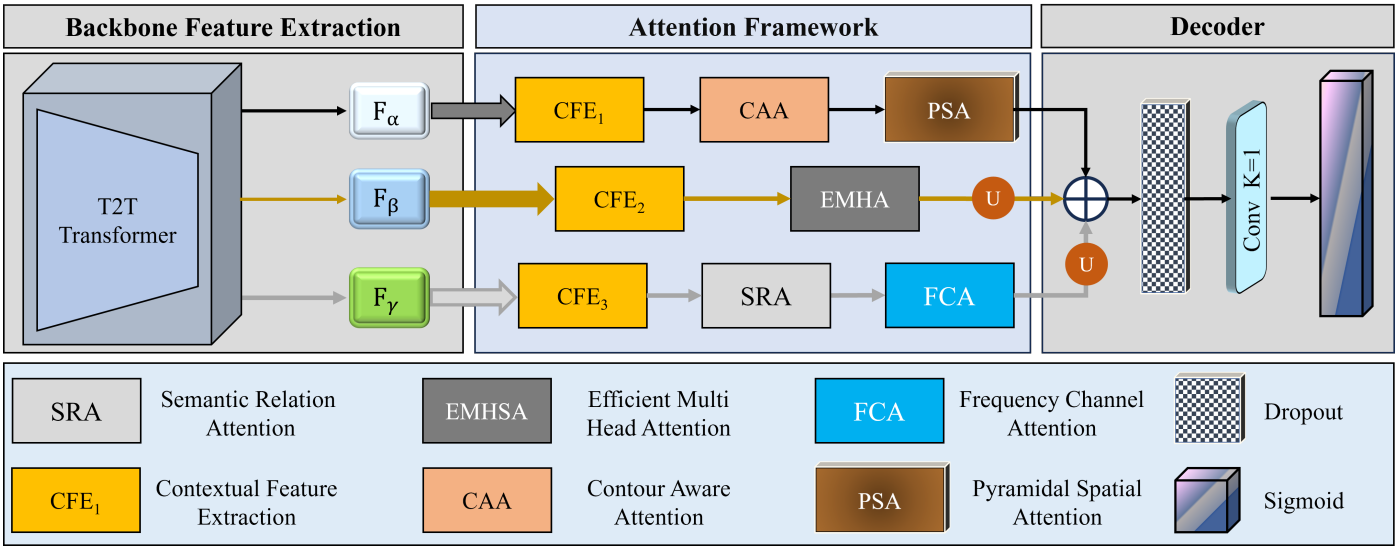

Figure 1. Overall architecture of the proposed MAFNet framework. The system consists of three main components: (1) T2T Transformer backbone extracting hierarchical features at three scales, (2) Attention framework with six specialized modules, and (3) Parallel multi-path decoder fusing attention-enhanced features for saliency map generation.

accuracy [15], while relational reasoning networks incorporate parallel multi-scale attention for semantic comprehension [26]. Deep supervision strategies within specialized architectures enable an effective combination of multi-scale features through nonlinear transformations [27]. Contemporary methods have introduced frequency-domain processing for multi-scale extraction [28] and dual-path architectures that combine convolutional networks with lightweight vision transformers for comprehensive feature representation [29].

2.3 Attention Mechanisms and Vision Transformers

Attention mechanisms have become fundamental components in modern SOD architectures, enabling selective focus on relevant features while suppressing background interference [30]. Progressive attention frameworks integrate multi-scale contextual features to enhance detection performance [21]. Multi-scale attention-guided modules effectively capture salient object scales via intelligent feature-weighting mechanisms [31]. Specialized attention architectures have been developed for specific applications, incorporating position enhancement and detail refinement through semantic and contextual attention modules [32]. Hybrid attention mechanisms facilitate effective multi-scale feature fusion when integrated with deep residual networks [33].

Vision transformers have revolutionized image analysis by treating visual data as sequential information, where image patches are processed as discrete tokens to model global dependencies. The Vision Transformer pioneered this paradigm by segmenting images into patches for transformer-based processing [34]. Subsequent developments improved efficiency through knowledge distillation strategies [35] and adapted transformers for dense prediction tasks using pyramid structures [36]. The Tokens-to-Token module enhanced local structure modeling and multi-scale feature extraction capabilities [37]. Recent SOD approaches have successfully integrated transformers with traditional CNN architectures. Global-to-local processing paradigms employ transformer backbones to enhance feature extraction and contextual interaction [38]. Asymmetric bilateral architectures fuse transformer and CNN features for comprehensive feature learning [39], while hierarchical transformer designs incorporate cross-modal interaction modules during decoding stages [40]. Advanced frameworks combine vision transformers with highly transformed decoders to preserve spatial detail [41]. These developments demonstrate the effectiveness of transformer-based architectures in capturing long-range dependencies essential for accurate SOD.

3 Proposed Methodology

This section presents a comprehensive description of our proposed attention-driven framework for efficient SOD. The overall architecture, illustrated in Figure 1, consists of three primary components: (1) a Tokens-to-Token (T2T) Transformer[37] backbone for hierarchical feature extraction, (2) a multi-level

attention framework for feature refinement and enhancement, and (3) a parallel multi-path decoder for multi-scale feature integration and saliency map generation.

3.1 Overall Architecture

Given an input RGB image $I \in \mathbb{R}^{H\times W\times 3}$ where H and W represent the height and width respectively, our network generates a saliency probability map $S \in \mathbb{R}^{H\times W}$ that identifies the most visually prominent objects. The architecture processes features through three main stages: backbone feature extraction, attention-based feature refinement, and multi-scale feature fusion. The T2T Transformer backbone extracts multi-scale features at three hierarchical levels, which are then processed through specialized attention modules tailored to their semantic characteristics. Finally, the decoder directly integrates these refined features from parallel attention-enhanced pathways to produce the final saliency prediction.

3.2 Backbone Feature Extraction

We employ the Tokens-to-Token Vision Transformer (T2T-ViT-14) as our feature-extraction backbone because it captures both local structural information and global contextual dependencies. Unlike conventional Vision Transformers that directly split images into fixed-size patches, T2T-ViT progressively tokenizes the image through a layered structure, enabling better modeling of local structures and inter-token relationships. For an input image of size $224 \times 224 \times 3$, the T2T-ViT-14 extracts hierarchical features at three distinct scales:

- $F_\alpha \in \mathbb{R}^{56\times 56\times 64}$: Shallow-level features with 64 channels capturing fine-grained spatial details and low-level patterns such as edges, textures, and color information. These features preserve high spatial resolution crucial for accurate boundary delineation.
- $F_\beta \in \mathbb{R}^{28\times 28\times 64}$: Mid-level features with 64 channels representing intermediate semantic concepts with moderate spatial resolution. These features balance spatial detail with semantic understanding, capturing object parts and local structures.
- $F_\gamma \in \mathbb{R}^{14\times 14\times 384}$: Deep-level features with 384 channels encoding high-level semantic information with rich contextual understanding. Despite lower spatial resolution, these features possess strong discriminative power for identifying salient regions.

The progressive reduction in spatial resolution (from 56×56 to 14×14), coupled with significantly increased channel dimensions at the deepest level (384 channels) and semantic richness, enables comprehensive scene understanding across multiple scales. This hierarchical feature extraction strategy provides a robust foundation for subsequent attention-based refinement and multi-scale feature integration.

3.3 Attention Framework

The extracted hierarchical features are processed through a sophisticated attention framework comprising six specialized modules, each designed to address specific aspects of feature refinement and enhancement.

3.3.1 Contextual Feature Extraction (CFE)

The CFE modules (CFE_1, CFE_2, CFE_3) refine features at different semantic levels by capturing multi-scale contextual information. Each CFE module processes its corresponding feature map through three parallel convolution branches with varying receptive fields to extract contextual information at different scales:

$$F_{\text{CFE}_i} = \text{Conv}_{1\times 1}(\text{Concat}[B_1(F_i), B_2(F_i), B_3(F_i)]) \tag{1}$$

where $F_i \in \{F_\alpha, F_\beta, F_\gamma\}$ represents the input feature from the backbone. Each branch B_j employs convolutions with different kernel sizes but produces a unified number of 64 output channels:

- $B_1(F_i)$: Applies 3×3 convolution → outputs 64 channels for local contextual patterns
- $B_2(F_i)$: Applies 5×5 convolution → outputs 64 channels for medium-range dependencies
- $B_3(F_i)$: Applies 7×7 convolution → outputs 64 channels for broader contextual relationships

The outputs from these three branches are concatenated along the channel dimension using Concat[·], resulting in 192 channels (64×3). A subsequent 1×1 convolution then fuses these concatenated features and reduces the channel dimensions to 64, producing the refined output $F_{\text{CFE}_i} \in \mathbb{R}^{H\times W\times 64}$. This multi-branch design enables the module to simultaneously capture contextual information at multiple scales, enriching feature representation and improving the network's ability to handle objects of varying sizes.

3.3.2 *Contour Aware Attention (CAA)*

Following CFE_1, the CAA module enhances boundary-related features by explicitly modeling object contours, which is crucial for accurate segmentation and precise delineation of salient object boundaries. Given the refined feature $F_{\text{CFE}_1} \in \mathbb{R}^{56\times56\times64}$, the CAA module first extracts edge information to identify potential object boundaries:

$$E = \text{EdgeDetect}(F_{\text{CFE}_1}) \tag{2}$$

where $\text{EdgeDetect}(\cdot)$ applies edge detection operations to extract boundary-related features. This operation uses convolutional edge-detection kernels that respond strongly to intensity gradients and structural discontinuities in the feature maps, thereby highlighting regions corresponding to object boundaries and contours. The extracted edge features E are then processed through a 3×3 convolution layer to refine and enhance the boundary representations:

$$E_{\text{refined}} = \text{Conv}_{3\times3}(E) \tag{3}$$

Subsequently, a sigmoid activation function $\sigma(\cdot)$ is applied to generate normalized contour-aware features in the range $[0, 1]$, indicating the likelihood of each spatial location being part of an object boundary:

$$F_{\text{CAA}} = \sigma(E_{\text{refined}}) \tag{4}$$

This mechanism emphasizes features along object boundaries while suppressing less relevant background regions, effectively guiding the network to focus on boundary-critical information. By explicitly modeling contours, the CAA module significantly improves segmentation precision, particularly in challenging scenarios where salient objects have complex or subtle boundaries. The output F_{CAA} retains the spatial dimensions $56 \times 56 \times 64$ while being enriched with enhanced boundary awareness.

3.3.3 *Pyramidal Spatial Attention (PSA)*

The PSA module processes the contour-aware features through a pyramidal spatial attention mechanism to capture multi-scale spatial dependencies and enhance the network's ability to handle objects of varying sizes. Given the input features $F_{\text{CAA}} \in \mathbb{R}^{56\times56\times64}$, the PSA module constructs a spatial pyramid by processing features at multiple scales through parallel branches. The pyramidal structure consists of multiple pooling operations with different kernel sizes to create feature representations at various spatial scales:

$$P_k = \text{Pool}_{k\times k}(F_{\text{CAA}}), \quad k \in \{1, 2, 4, 8\} \tag{5}$$

where P_k represents the pooled features at scale k, capturing spatial context at different granularities. The scale $k = 1$ preserves the original resolution, while larger scales capture increasingly global spatial context. Each pooled feature is then processed through a convolution layer and upsampled back to the original spatial dimensions:

$$P_k^{\uparrow} = \text{Upsample}(\text{Conv}_{1\times1}(P_k)), \quad \forall k \tag{6}$$

The multi-scale features are aggregated through concatenation and fused through convolution layers to generate spatially enhanced features:

$$F_{\text{PSA}} = \text{Conv}_{3\times3}(\text{Concat}[P_1^{\uparrow}, P_2^{\uparrow}, P_4^{\uparrow}, P_8^{\uparrow}]) \tag{7}$$

This pyramidal approach enables the network to effectively capture both local fine-grained details and global spatial context, making it robust to scale variations in salient objects. The output F_{PSA} preserves the spatial dimensions $56 \times 56 \times 64$ and is enriched with multi-scale spatial information that emphasizes salient regions across scales.

3.3.4 *Efficient Multi-Head Attention (EMHA)*

The EMHA module enhances the mid-level features F_{CFE_2} by computing multi-head self-attention efficiently:

$$F_{\text{EMHA}} = \text{MHA}(Q, K, V) + F_{\text{CFE}_2} \tag{8}$$

where Q, K, V are query, key, and value projections of F_{CFE_2}, and $\text{MHA}(\cdot)$ denotes the multi-head attention operation. This module captures long-range dependencies within the feature space, enabling better understanding of object-context relationships.

3.3.5 *Semantic Relation Attention (SRA)*

Non-Local Neural Networks [42] inspire the SRA module and process deep-level features $F_{\text{CFE}_3} \in \mathbb{R}^{14\times14\times64}$ to model semantic relationships between spatial regions via non-local operations. Unlike local convolutions that capture information within a limited receptive field, the non-local operation computes a response at a position as a weighted sum of features from all positions, enabling the network to capture long-range dependencies regardless of spatial distance. Given the input features F_{CFE_3}, the non-local operation first reshapes the features from $\mathbb{R}^{H\times W\times C}$ to $\mathbb{R}^{N\times C}$ where $N = H \times W$ represents the total number of spatial locations. Three embedding functions are then

applied to generate query (Q), key (K), and value (V) representations:

$$Q = F_{\text{CFE}_3} W_Q, \quad K = F_{\text{CFE}_3} W_K, \quad V = F_{\text{CFE}_3} W_V \tag{9}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{C \times C'}$ are learnable transformation matrices, and C' is the reduced channel dimension for computational efficiency. The pairwise relationships between all spatial locations are computed through the dot product of queries and keys:

$$f(x_i, x_j) = e^{Q_i \cdot K_j} \tag{10}$$

where Q_i and K_j represent the query at position i and key at position j, respectively. This measures the semantic similarity or affinity between any two positions in the feature map. The pairwise affinities are then normalized using softmax to obtain attention weights:

$$A_{ij} = \frac{f(x_i, x_j)}{\sum_{\forall j} f(x_i, x_j)} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{11}$$

where A_{ij} represents the attention weight from position j to position i, and $\sqrt{d_k}$ is a scaling factor that stabilizes gradients during training. These attention weights capture the semantic relationships between all pairs of spatial locations, allowing the network to understand which regions are semantically related regardless of their spatial proximity. The non-local response is then computed by aggregating value features weighted by the attention map:

$$y_i = \sum_{\forall j} A_{ij} V_j \tag{12}$$

Finally, a residual connection combines the non-local response with the original input features:

$$F_{\text{SRA}} = F_{\text{CFE}_3} + W_z \cdot y \tag{13}$$

where W_z is a learnable projection matrix that transforms the aggregated features back to the original channel dimension, and y represents the non-local response reshaped to $\mathbb{R}^{H \times W \times C'}$. This residual formulation ensures that the module can be inserted into any pre-trained network without breaking its initial behavior. The non-local operation enables the network to capture global contextual relationships and model semantic dependencies between distant regions, which is essential for identifying salient objects based on holistic scene understanding. The output F_{SRA} preserves the dimensions $14 \times 14 \times 64$ while incorporating non-local semantic context that captures relationships across the entire feature map.

3.3.6 *Frequency Channel Attention (FCA)*

The FCA module enhances features by operating in the frequency domain, leveraging spectral information to emphasize important channels for saliency detection selectively. Given the semantically enhanced features $F_{\text{SRA}} \in \mathbb{R}^{14 \times 14 \times 64}$, the FCA module first transforms the spatial features into the frequency domain using the Fast Fourier Transform (FFT):

$$F_{\text{freq}} = \text{FFT}(F_{\text{SRA}}) \tag{14}$$

where $\text{FFT}(\cdot)$ converts the spatial representation into its frequency components, capturing both low-frequency (global structure) and high-frequency (fine details) information. The frequency representation enables the network to analyze features from a complementary perspective, as different frequency components often correspond to different semantic or structural properties. To generate channel attention weights, the frequency features are processed through global average pooling across the spatial dimensions, followed by a channel-wise fully connected layer:

$$A_{\text{channel}} = \sigma(\text{FC}(\text{GAP}(|F_{\text{freq}}|))) \tag{15}$$

where $|\cdot|$ computes the magnitude of the complex-valued frequency features, $\text{GAP}(\cdot)$ performs global average pooling, $\text{FC}(\cdot)$ denotes a fully connected layer that learns channel importance, and $\sigma(\cdot)$ is the sigmoid activation function that normalizes the attention weights to $[0, 1]$. These channel attention weights $A_{\text{channel}} \in \mathbb{R}^{1 \times 1 \times 64}$ indicate the relative importance of each feature channel for saliency detection. The frequency features are then transformed back to the spatial domain and modulated by the channel attention weights:

$$F_{\text{FCA}} = \text{IFFT}(F_{\text{freq}}) \odot A_{\text{channel}} \tag{16}$$

where $\text{IFFT}(\cdot)$ denotes the Inverse Fast Fourier Transform that converts frequency features back to the spatial domain, and $\odot$ represents channel-wise multiplication that applies the attention weights. By operating in the frequency domain, the FCA module can effectively suppress noise and emphasize frequency components that are most discriminative for identifying salient regions. The output F_{FCA} maintains the spatial dimensions $14 \times 14 \times 64$ while being enriched with frequency-aware channel attention that highlights the most informative channels for saliency prediction.

3.4 Parallel Multi-Path Decoder

The decoder fuses refined multi-scale features from three parallel attention-enhanced pathways

via direct integration. Rather than employing stage-by-stage refinement, our decoder concurrently processes features across all semantic levels, enabling efficient feature aggregation while preserving complementary information across scales. The three attention-enhanced feature streams (F_{PSA}, F_{EMHA}, F_{FCA}) are initially spatially aligned to uniform resolution via upsampling operations:

$$\begin{aligned} F_{\text{PSA}}^{\uparrow} &= \text{Upsample}(F_{\text{PSA}}, \text{size} = H_{\text{target}} \times W_{\text{target}}) \\ F_{\text{EMHA}}^{\uparrow} &= \text{Upsample}(F_{\text{EMHA}}, \text{size} = H_{\text{target}} \times W_{\text{target}}) \\ F_{\text{FCA}}^{\uparrow} &= \text{Upsample}(F_{\text{FCA}}, \text{size} = H_{\text{target}} \times W_{\text{target}}) \end{aligned} \tag{17}$$

where H_{target} and W_{target} specify the target spatial dimensions for fusion. The aligned features are subsequently merged through element-wise addition to aggregate complementary information across semantic levels:

$$F_{\text{fused}} = F_{\text{PSA}}^{\uparrow} \oplus F_{\text{EMHA}}^{\uparrow} \oplus F_{\text{FCA}}^{\uparrow} \tag{18}$$

where $\oplus$ represents element-wise addition. This parallel fusion mechanism enables simultaneous exploitation of fine-grained spatial details from F_{PSA}, intermediate semantic representations from F_{EMHA}, and high-level contextual information from F_{FCA}. The fused features undergo further refinement via convolutional layers with dropout regularization to strengthen feature representation and mitigate overfitting:

$$F_{\text{decode}} = \text{Dropout}(\text{Conv}_{3\times3}(\text{ReLU}(\text{Conv}_{3\times3}(F_{\text{fused}})))) \tag{19}$$

Subsequently, a 1×1 convolution with sigmoid activation produces the saliency probability map:

$$S = \sigma(\text{Conv}_{1\times1}(F_{\text{decode}})) \tag{20}$$

This parallel multi-path fusion strategy ensures effective integration of multi-scale features while preserving computational efficiency, yielding accurate, spatially refined saliency predictions that balance boundary precision with semantic comprehension.

3.5 Loss Function

We utilize a hybrid loss function combining binary cross-entropy (BCE) and structural similarity (SSIM) losses for network optimization:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{BCE}}\mathcal{L}_{\text{BCE}} + \lambda_{\text{SSIM}}\mathcal{L}_{\text{SSIM}} \tag{21}$$

where λ_{BCE} and λ_{SSIM} denote weighting coefficients. The BCE loss enforces pixel-level accuracy:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[G_i \log(S_i) + (1-G_i)\log(1-S_i)] \tag{22}$$

where G represents the ground truth saliency map, S denotes the predicted map, and N indicates the total pixel count. The SSIM loss maintains structural consistency:

$$\mathcal{L}_{\text{SSIM}} = 1 - \text{SSIM}(S, G) \tag{23}$$

This hybrid loss formulation promotes both precise pixel-wise predictions and the preservation of structural similarity relative to ground-truth annotations.

4 Results and Discussion

4.1 Experimental Setup

Datasets We trained our network on the DUTS-TR dataset [43], which comprises 10,553 diverse images with varying levels of complexity and salient object types. For a comprehensive evaluation, we tested our network on six widely used benchmark datasets. The ECSSD dataset [44] comprises 1,000 images with intricate, semantically rich structures, thereby challenging models to accurately identify complex salient objects. HKU-IS [18] contains 4,447 challenging images characterized by disconnected salient objects, boundary overlapping, and low color contrast, testing the model's ability to handle complex configurations. DUT-OMRON [4] provides 5,168 high-quality images featuring one or more salient objects against complex backgrounds, assessing the model's performance in intricate scenes. PASCAL-S [45] includes 850 images emphasizing distinct salient objects, offering diverse instances for evaluating performance across different object types and contexts. DUTS-TE [43], a testing subset of DUTS with 5,019 images, features complex backgrounds and varying salient object types. Finally, SOD [46] provides 300 natural images with diverse salient object types, evaluating the model's robustness across various scenarios. This comprehensive evaluation across multiple datasets ensures a thorough assessment of our model's generalization capability and effectiveness in handling diverse saliency detection challenges.

Evaluation Metrics We employ five standard evaluation metrics to comprehensively assess performance. The S-measure (S_m) assesses structural similarity by comparing predicted saliency maps with ground-truth saliency maps through region-aware and object-aware structural analyses. The F-measure (F_β) calculates the weighted harmonic mean of precision and recall, where $\beta^2 = 0.3$ places greater emphasis on precision. The E-measure (E_m) evaluates

both local pixel-level accuracy and global image statistics by combining pixel values with image-level mean information. Mean Absolute Error ($\mathcal{M}$) quantifies the average pixel-wise absolute difference between predictions and ground truth. For F_β and E_m, we report the maximum values obtained across all threshold levels. Better performance is indicated by higher values of S_m, F_β, and E_m, along with lower values of $\mathcal{M}$.

Implementation Details Our network is implemented in PyTorch and trained on an NVIDIA RTX 3090 GPU with 24GB of memory. All input images are resized to 224×224 pixels during both training and testing phases. We employ the Adam optimizer with an initial learning rate of $1e-4$ and weight decay of $5e-4$ for network optimization. The model is trained for 100 epochs with a batch size of 8. We set the loss function weighting coefficients to $\lambda_{\text{BCE}} = 1.0$ and $\lambda_{\text{SSIM}} = 1.0$, giving equal importance to both binary cross-entropy and structural similarity losses. For data augmentation, we apply random horizontal flipping and random rotation to enhance model robustness and prevent overfitting. The learning rate is reduced by a factor of 0.1 at epochs 60 and 80, using a step-decay schedule. The T2T-ViT-14 backbone is initialized with weights pre-trained on ImageNet, while all attention modules are randomly initialized. During inference, we directly feed the input images into the network without any post-processing operations to generate the final saliency maps.

4.2 Comparison with State-of-the-Art Methods

We conduct comprehensive quantitative comparisons with SOTA RGB methods, including PiCANet[47], R3Net[48], CPD[49], BASNet[11], EGNet[50], GateNet[51], U2Net[52], MINet[53], and VST[9]. For qualitative visualization, we additionally include CPD-R (a residual-refinement variant of CPD), F3Net [54], and LDF [55], which are representative boundary-aware SOD methods commonly used for visual comparison in recent literature. These methods represent diverse architectural paradigms, including attention-based mechanisms and transformer-based architectures, providing a comprehensive benchmark for evaluating the proposed framework.

4.2.1 Quantitative Analysis

Tables 1 and 2 present comprehensive quantitative comparisons of our proposed method against nine state-of-the-art RGB SOD methods across six benchmark datasets. The results demonstrate that our approach achieves superior or competitive performance across all evaluation metrics. On the ECSSD dataset, our method achieves the highest performance with $F_\beta = 0.949$, $\mathcal{M} = 0.031$, and $S_m = 0.936$, surpassing strong baselines, including MINet and VST. The E_m score of 0.964 matches VST's performance, indicating excellent structural preservation. Specifically, our method improves F_β over MINet and S_m over VST, while maintaining the lowest MAE of 0.031. For PASCAL-S, our approach achieves the best results across all metrics $F_\beta = 0.887$, $\mathcal{M} = 0.065$, $S_m = 0.878$, $E_m = 0.911$, showing substantial improvements over competing methods. This represents significant gains over the second-best method, MINet, with enhancements in F_β and S_m, indicating our model's robustness in handling diverse salient object types and contextual variations. On the SOD dataset, we obtain $F_\beta = 0.865$ and $S_m = 0.857$, representing the highest scores among all compared methods, while the E_m score of 0.904 demonstrates superior edge-aware detection capability.

Results on the DUTS-TE dataset further validate our method's effectiveness, with our approach achieving the best performance across all metrics $F_\beta = 0.898$, $\mathcal{M} = 0.034$, $S_m = 0.905$, $E_m = 0.941$. Compared with the second-best methods, we achieve improvement in F_β over MINet and in S_m over VST, demonstrating superior generalization on this challenging test set with complex backgrounds and diverse object types. On HKU-IS, which contains images with disconnected salient objects and boundary overlap, our method achieves $F_\beta = 0.943$, $\mathcal{M} = 0.026$, $S_m = 0.931$, and $E_m = 0.963$, outperforming all competing approaches. This highlights the effectiveness of our Contour Aware Attention and Pyramidal Spatial Attention modules in handling complex object configurations. Notably, we achieve improvements in F_β over MINet and in S_m over VST, while achieving the lowest MAE among all methods. For DUT-OMRON, our method achieves consistent improvements with $F_\beta = 0.813$, $\mathcal{M} = 0.054$, $S_m = 0.855$, and $E_m = 0.891$, demonstrating robust performance in detecting salient objects against complex backgrounds. While the performance gains over the strongest baselines are sometimes modest in absolute terms (e.g., 0.005 in F_β on ECSSD over VST), they are consistent across all six benchmark datasets and all four evaluation metrics, suggesting that the improvements stem from systematic architectural advantages rather than dataset-specific variance. Future work will incorporate multiple training runs with statistical significance testing to further substantiate these findings.

Table 1. Quantitative comparison with state-of-the-art RGB salient object detection methods on ECSSD, PASCAL-S, and SOD datasets. Best results are highlighted in **bold**. ↑ indicates higher is better, ↓ indicates lower is better. "-" indicates the metric was not reported in the original publication.

Method	ECSSD				PASCAL-S				SOD			
	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$E_m \uparrow$
PiCANet	0.919	0.046	0.914	0.947	0.836	0.078	0.848	0.896	0.854	0.103	0.789	0.796
R3Net	0.926	0.040	0.910	0.949	0.800	0.092	0.807	0.853	0.850	0.125	0.759	0.796
CPD	0.939	0.037	0.918	0.944	0.836	0.072	0.845	0.888	0.860	0.112	0.767	0.778
BASNet	0.931	0.034	0.925	0.947	0.835	0.076	0.838	0.886	0.808	0.102	0.793	0.822
EGNet	0.943	0.041	0.918	0.946	0.869	0.074	0.852	0.877	0.890	0.097	0.807	0.842
GateNet	0.928	0.041	0.917	0.949	0.850	0.068	0.858	0.901	0.835	0.079	0.827	0.877
U2Net	0.941	0.033	0.928	0.957	0.832	0.074	0.845	0.883	0.861	0.108	0.786	-
MINet	0.947	0.033	0.925	0.953	0.882	0.064	0.857	0.899	0.835	0.074	0.830	0.878
VST	0.944	0.034	0.932	0.964	0.850	0.067	0.873	0.900	0.866	0.065	0.854	0.902
Ours	**0.949**	**0.031**	**0.936**	**0.964**	**0.887**	**0.065**	**0.878**	**0.911**	**0.865**	**0.073**	**0.857**	**0.904**

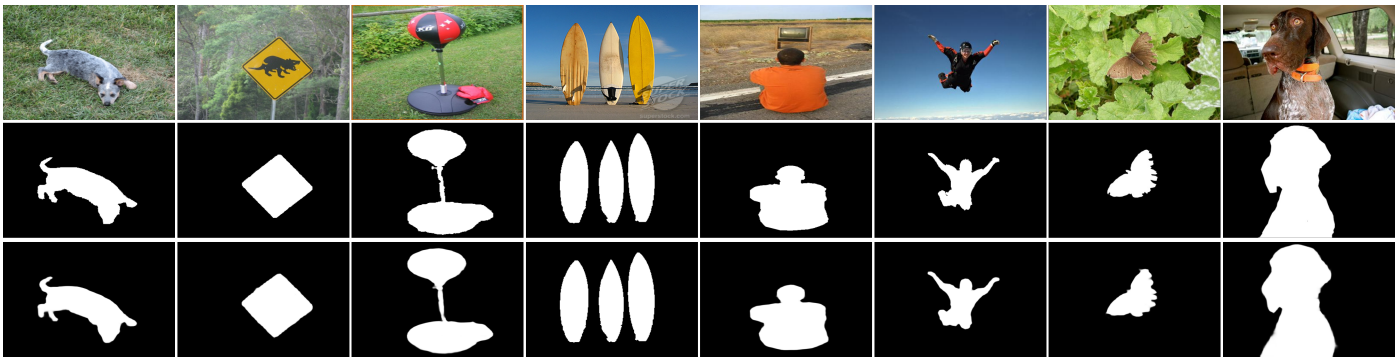

Figure 2. Qualitative evaluation of MAFNet on challenging scenarios from six benchmark datasets. Top row: Original RGB input images, Middle row: Ground truth, and Bottom row: Predicted saliency maps generated by our method. Results demonstrate MAFNet's effectiveness in handling objects of varying sizes, maintaining sharp boundary delineation, and suppressing complex background interference.

Qualitative evaluation of MAFNet on challenging scenarios from six benchmark datasets. Each column presents a different test case demonstrating the model's robustness across diverse conditions. Top row: Original RGB input images, Middle row: Ground truth binary saliency annotations, and Bottom row: Predicted saliency maps generated by our method. The results demonstrate MAFNet's effectiveness in handling objects of varying sizes, maintaining sharp boundary delineation, achieving complete object coverage, and suppressing complex background interference across different datasets.

4.2.2 *Qualitative Analysis*

Figure 2 presents visual results of our proposed method across diverse challenging scenarios from benchmark datasets. Our approach demonstrates robust performance in detecting salient objects of varying sizes, shapes, and complexities. The technique effectively handles objects such as signs and surfboards, maintaining sharp edges and complete object coverage. For challenging cases, our model successfully captures the entire salient region. These results validate the effectiveness of our Contour Aware Attention module in preserving object boundaries and the T2T Transformer backbone in capturing global contextual information.

Figure 3 provides visual comparisons with state-of-the-art methods, including BASNet, CPD-R, F3Net, LDF, and PiCANet, across five challenging scenarios. Our method produces cleaner saliency maps with more complete object coverage than other approaches, which often yield fragmented or incomplete detections. Our approach yields more uniform saliency predictions with greater structural

Table 2. Quantitative comparison with state-of-the-art RGB salient object detection methods on DUTS-TE, HKU-IS, and DUT-OMRON datasets.

Method	DUTS-TE				HKU-IS				DUT-OMRON			
	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$E_m \uparrow$
PiCANet	0.827	0.054	0.861	0.907	0.908	0.042	0.906	0.950	0.759	0.068	0.826	0.867
R3Net	0.800	0.058	0.831	0.881	0.904	0.036	0.895	0.945	0.760	0.063	0.817	0.857
CPD	0.839	0.043	0.867	0.912	0.909	0.033	0.904	0.948	0.747	0.057	0.818	0.856
BASNet	0.839	0.048	0.866	0.903	0.919	0.032	0.909	0.952	0.779	0.057	0.836	0.872
EGNet	0.893	0.039	0.875	0.904	0.937	0.031	0.918	0.956	0.842	0.052	0.818	0.874
GateNet	0.847	0.045	0.870	0.916	0.916	0.036	0.910	0.952	0.754	0.061	0.821	0.858
U2Net	0.849	0.045	0.874	0.911	0.924	0.031	0.916	0.954	0.793	0.055	0.847	0.880
MINet	0.884	0.037	0.884	0.917	0.935	0.028	0.920	0.961	0.810	0.055	0.833	0.873
VST	0.877	0.037	0.896	0.939	0.937	0.030	0.928	0.968	0.800	0.058	0.850	0.888
Ours	**0.898**	**0.034**	**0.905**	**0.941**	**0.943**	**0.026**	**0.931**	**0.963**	**0.813**	**0.054**	**0.855**	**0.891**

Table 3. Component-wise ablation study. CFE: Contextual Feature Extraction, CAA: Contour Aware Attention, PSA: Pyramidal Spatial Attention, EMHA: Efficient Multi-Head Attention, SRA: Semantic Relation Attention, FCA: Frequency Channel Attention.

Components							ECSSD				PASCAL-S				DUTS-TE			
CFE	CAA	PSA	EMHA	SRA	FCA	Baseline	F_β	$\mathcal{M}$	S_m	E_m	F_β	$\mathcal{M}$	S_m	E_m	F_β	$\mathcal{M}$	S_m	E_m
						✓	0.928	0.039	0.923	0.951	0.871	0.073	0.865	0.898	0.879	0.041	0.893	0.928
✓							0.937	0.036	0.928	0.955	0.876	0.070	0.870	0.902	0.885	0.039	0.897	0.933
✓	✓						0.941	0.034	0.931	0.958	0.879	0.068	0.872	0.904	0.889	0.037	0.900	0.936
✓	✓	✓					0.944	0.033	0.933	0.960	0.882	0.067	0.875	0.907	0.892	0.036	0.902	0.938
✓	✓	✓	✓				0.946	0.032	0.934	0.962	0.884	0.066	0.876	0.909	0.895	0.035	0.903	0.939
✓	✓	✓	✓	✓			0.948	0.032	0.935	0.963	0.886	0.066	0.877	0.910	0.897	0.035	0.904	0.940
✓	✓	✓	✓	✓	✓		**0.949**	**0.031**	**0.936**	**0.964**	**0.887**	**0.065**	**0.878**	**0.911**	**0.898**	**0.034**	**0.905**	**0.941**

consistency, whereas other methods produce patchy or incomplete results. Our method yields the cleanest results, with precise boundary preservation and complete object coverage, whereas other methods either miss parts or introduce background interference. These qualitative comparisons confirm that our specialized multi-level attention framework effectively addresses the limitations of existing methods, producing more accurate and visually coherent saliency predictions across diverse challenging scenarios.

4.3 Ablation Study

We conduct two ablation studies to validate our design choices: (1) component-wise analysis to evaluate the contribution of each attention module, and (2) backbone comparison to justify our choice of T2T Transformer as the feature extractor.

4.3.1 Component-wise Analysis

Table 3 presents the progressive addition of attention modules to evaluate their individual contributions. The baseline model comprises the T2T Transformer backbone and a simple decoder. Adding CFE modules for multi-scale contextual extraction improves F_β from 0.928 to 0.937 on ECSSD, an absolute gain of 0.9 percentage points, confirming the benefit of aggregating contextual information across multiple receptive fields. The CAA module further enhances boundary preservation, reducing MAE by 0.004 on the ECSSD dataset. PSA brings a further improvement in F_β from 0.879 to 0.882 on PASCAL-S, indicating that the pyramidal pooling structure improves robustness to object scale variation beyond what CFE alone provides. EMHA and SRA modules enhance semantic understanding via attention mechanisms, whereas FCA provides the final boost

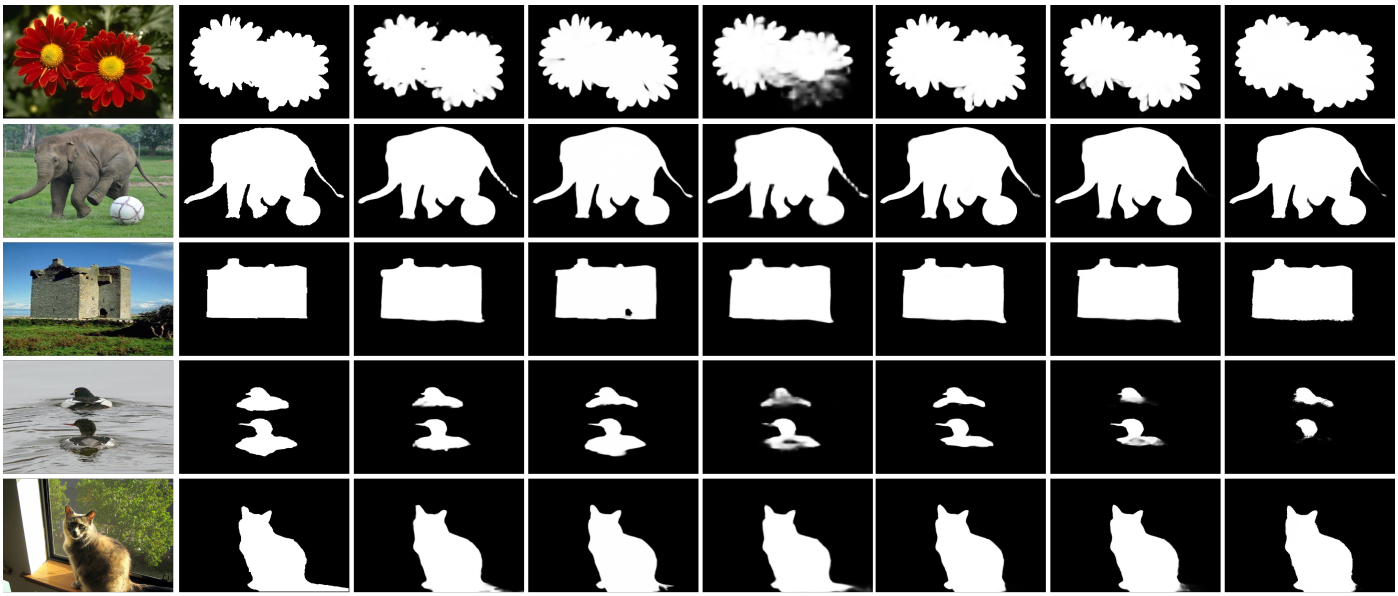

Figure 3. Visual comparison of our method with state-of-the-art approaches. From left to right: input RGB image, ground truth (GT), and predictions from our method, BASNet, CPD-R, F3Net, LDF, and PiCANet.

Table 4. Backbone architecture comparison. All models use the same attention framework and decoder.

Backbone	ECSSD				HKU-IS				DUT-OMRON			
	F_β	$\mathcal{M}$	S_m	E_m	F_β	$\mathcal{M}$	S_m	E_m	F_β	$\mathcal{M}$	S_m	E_m
ResNet-50	0.931	0.037	0.921	0.952	0.928	0.033	0.918	0.955	0.795	0.061	0.842	0.881
PVT-Small	0.941	0.034	0.930	0.958	0.936	0.029	0.925	0.959	0.804	0.057	0.849	0.886
Swin-Tiny	0.945	0.032	0.933	0.961	0.940	0.028	0.928	0.961	0.808	0.056	0.851	0.888
T2T-ViT-14	**0.949**	**0.031**	**0.936**	**0.964**	**0.943**	**0.026**	**0.931**	**0.963**	**0.813**	**0.054**	**0.855**	**0.891**

via frequency-domain channel attention. Each module contributes meaningful improvements, and their combination achieves the best performance across all metrics.

4.3.2 *Backbone Comparison*

Table 4 compares different backbone architectures to validate our choice of T2T-ViT-14. We evaluate ResNet-50, PVT-Small, Swin-Transformer-Tiny, and T2T-ViT-14, keeping the attention framework and decoder identical. ResNet-50, while efficient, struggles to capture global context due to limited receptive fields. PVT-Small demonstrates improved performance through a hierarchical transformer design but lacks a progressive tokenization strategy. Swin-Transformer-Tiny achieves competitive results with shifted-window attention but incurs a higher computational cost. T2T-ViT-14 achieves the best balance between performance and efficiency, with superior F_β and S_m scores across all datasets. The progressive tokenization in T2T Transformer effectively models local structures while maintaining global context awareness, justifying our architectural choice.

5 Conclusion

In this paper, we presented MAFNet, a novel multi-level attention fusion network for salient object detection. Our approach introduces a comprehensive attention framework with six specialized modules tailored to different semantic levels: CFE for multi-scale context extraction, CAA for boundary preservation, PSA for hierarchical spatial reasoning, EMHA for semantic enhancement, SRA for global context modeling, and FCA for frequency-domain feature enhancement. By integrating these specialized attention mechanisms with a T2T Transformer backbone and parallel multi-path decoder, MAFNet effectively addresses key limitations in existing SOD methods. Extensive experiments on six benchmark datasets demonstrate that MAFNet achieves state-of-the-art performance. Ablation

studies confirm that each attention module contributes meaningfully to overall performance. Future work will explore extending this framework to related tasks, such as RGB-D salient object detection and camouflaged object detection, and develop lightweight variants for mobile deployment.

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 Claude was used to improve the readability and language quality of the manuscript. The authors have carefully reviewed and edited the AI-assisted output and take full responsibility for the content of the manuscript.

Ethical Approval and Consent to Participate

Not applicable. This study did not involve human participants, animal subjects, or clinical data. All experiments were conducted using publicly available benchmark datasets that do not contain personally identifiable information.

References

[1] Borji, A., Cheng, M. M., Jiang, H., & Li, J. (2015). Salient object detection: A benchmark. *IEEE transactions on image processing*, *24*(12), 5706-5722. [CrossRef]

[2] Wang, Y., Wang, R., Liu, J., Xu, R., Wang, T., Hou, F., ... & Lei, N. (2025). TFGNet: Frequency-guided saliency detection for complex scenes. *Applied Soft Computing*, *170*, 112685. [CrossRef]

[3] Borji, A., & Itti, L. (2012, June). Exploiting local and global patch rarities for saliency detection. In *2012 IEEE conference on computer vision and pattern recognition* (pp. 478-485). IEEE. [CrossRef]

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

[5] Huo, L., Jiao, L., Wang, S., & Yang, S. (2016). Object-level saliency detection with color attributes. *Pattern recognition*, *49*, 162-173. [CrossRef]

[6] Amjoud, A. B., & Amrouch, M. (2023). Object detection using deep learning, CNNs and vision transformers: A review. *IEEE Access*, *11*, 35479-35516. [CrossRef]

[7] Wei, L., Zhu, Z., Mi, Y., & Hu, W. (2025). PDNet: Pluralistic depth-aware network for RGB-D salient object detection. *Signal Processing*, 110271. [CrossRef]

[8] Wan, B., Zhou, X., Sun, Y., Wang, T., Lv, C., Wang, S., ... & Yan, C. (2023). MFFNet: Multi-modal feature fusion network for VDT salient object detection. *IEEE Transactions on Multimedia*, *26*, 2069-2081. [CrossRef]

[9] Liu, N., Zhang, N., Wan, K., Shao, L., & Han, J. (2021, October). Visual Saliency Transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 4702-4712). IEEE. [CrossRef]

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

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

[12] Usman, M. T., Khan, H., Rida, I., & Koo, J. (2025). Lightweight transformer-driven multi-scale trapezoidal attention network for saliency detection. *Engineering Applications of Artificial Intelligence*, *155*, 110917. [CrossRef]

[13] Chen, S., Tan, X., Wang, B., & Hu, X. (2018, September). Reverse Attention for Salient Object Detection. In *European Conference on Computer Vision* (pp. 236-252). Cham: Springer International Publishing. [CrossRef]

[14] Zhao, T., & Wu, X. (2019, June). Pyramid Feature Attention Network for Saliency Detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3080-3089). IEEE. [CrossRef]

[15] Zhang, J., Shi, Y., Zhang, Q., Cui, L., Chen, Y., & Yi, Y. (2022). Attention guided contextual feature fusion network for salient object detection. *Image and Vision Computing*, *117*, 104337. [CrossRef]

[16] Khan, H., Usman, M. T., Rida, I., & Koo, J. (2024). Attention enhanced machine instinctive vision with human-inspired saliency detection. *Image and Vision Computing*, *152*, 105308. [CrossRef]

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

[18] Li, G., & Yu, Y. (2015, June). Visual saliency based on multiscale deep features. In *2015 IEEE Conference on*

Computer Vision and Pattern Recognition (*CVPR*) (pp. 5455-5463). IEEE. [CrossRef]

[19] Liu, N., & Han, J. (2016, June). DHSNet: Deep Hierarchical Saliency Network for Salient Object Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*) (pp. 678-686). IEEE. [CrossRef]

[20] Wang, L., Wang, L., Lu, H., Zhang, P., & Ruan, X. (2016, September). Saliency detection with recurrent fully convolutional networks. In *European conference on computer vision* (pp. 825-841). Cham: Springer International Publishing. [CrossRef]

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

[22] Zeng, Y., Lu, H., Zhang, L., Feng, M., & Borji, A. (2018, June). Learning to Promote Saliency Detectors. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1644-1653). IEEE. [CrossRef]

[23] Jiang, Y., Yang, Z., Deng, L., & Zhou, T. (2025). Multi-Scale attention Coordination Network for remote sensing image salient object detection. *Optics & Laser Technology, 192*, 113751. [CrossRef]

[24] Islam, M. A., Kalash, M., & Bruce, N. D. (2018, June). Revisiting Salient Object Detection: Simultaneous Detection, Ranking, and Subitizing of Multiple Salient Objects. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7142-7150). IEEE. [CrossRef]

[25] Jia, S., & Bruce, N. D. (2019). Richer and deeper supervision network for salient object detection. *arXiv preprint arXiv:1901.02425*. [CrossRef]

[26] Cong, R., Zhang, Y., Fang, L., Li, J., Zhao, Y., & Kwong, S. (2021). RRNet: Relational reasoning network with parallel multiscale attention for salient object detection in optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing, 60*, 1-11. [CrossRef]

[27] Liu, Y., Cheng, M. M., Zhang, X. Y., Nie, G. Y., & Wang, M. (2021). DNA: Deeply supervised nonlinear aggregation for salient object detection. *IEEE Transactions on Cybernetics, 52*(7), 6131-6142. [CrossRef]

[28] Song, S., Jia, Z., Yang, J., & Kasabov, N. (2022). Salient detection via the fusion of background-based and multiscale frequency-domain features. *Information Sciences, 618*, 53-71. [CrossRef]

[29] Chen, Z., Lu, Y., Long, S., & Bai, J. (2024). Dual-path multi-branch feature residual network for salient object detection. *Engineering Applications of Artificial Intelligence, 133*, 108530. [CrossRef]

[30] Zhang, L., Li, X., Sun, Y., & Guo, H. (2025). Triple-attentions based salient object detector for strip steel surface defects. *Scientific Reports, 15*(1), 2537. [CrossRef]

[31] Noori, M., Mohammadi, S., Majelan, S. G., Bahri, A., & Havaei, M. (2020). DFNet: Discriminative feature extraction and integration network for salient object detection. *Engineering Applications of Artificial Intelligence, 89*, 103419. [CrossRef]

[32] Lin, Y., Sun, H., Liu, N., Bian, Y., Cen, J., & Zhou, H. (2022, September). Attention guided network for salient object detection in optical remote sensing images. In *International conference on artificial neural networks* (pp. 25-36). Cham: Springer International Publishing. [CrossRef]

[33] Li, H., Han, Y., Li, P., Li, X., & Shi, L. (2023). Hybrid attention mechanism and forward feedback unit for RGB-D salient object detection. *IEEE Access, 11*, 96068-96080. [CrossRef]

[34] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. [CrossRef]

[35] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021, July). Training data-efficient image transformers & distillation through attention. In *International conference on machine learning* (pp. 10347-10357). PMLR. https://proceedings.mlr.press/v139/touvron21a

[36] Wang, W., Xie, E., Li, X., Fan, D. P., Song, K., Liang, D., ... & Shao, L. (2021, October). Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In *2021 IEEE/CVF International Conference on Computer Vision* (*ICCV*) (pp. 548-558). IEEE. [CrossRef]

[37] Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z., ... & Yan, S. (2021, October). Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet. In *2021 IEEE/CVF International Conference on Computer Vision* (*ICCV*) (pp. 538-547). IEEE. [CrossRef]

[38] Li, G., Bai, Z., Liu, Z., Zhang, X., & Ling, H. (2023). Salient object detection in optical remote sensing images driven by transformer. *IEEE Transactions on Image Processing, 32*, 5257-5269. [CrossRef]

[39] Qiu, Y., Liu, Y., Zhang, L., Lu, H., & Xu, J. (2023). Boosting salient object detection with transformer-based asymmetric bilateral U-Net. *IEEE Transactions on Circuits and Systems for Video Technology, 34*(4), 2332-2345. [CrossRef]

[40] Zeng, C., Kwong, S., & Ip, H. (2023). Dual swin-transformer based mutual interactive network for RGB-D salient object detection. *Neurocomputing, 559*, 126779. [CrossRef]

[41] Ren, S., Zhao, N., Wen, Q., Han, G., & He, S. (2024). Unifying global-local representations in salient object detection with transformers. *IEEE Transactions on Emerging Topics in Computational Intelligence, 8*(4), 2870-2879. [CrossRef]

[42] Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7794-7803). [CrossRef]

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

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

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

[46] Movahedi, V., & Elder, J. H. (2010, June). Design and perceptual validation of performance measures for salient object segmentation. In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops* (pp. 49-56). IEEE. [CrossRef]

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

[48] Deng, Z., Hu, X., Zhu, L., Xu, X., Qin, J., Han, G., & Heng, P. A. (2018, July). R3net: Recurrent residual refinement network for saliency detection. In *Proceedings of the 27th international joint conference on artificial intelligence* (Vol. 684690). Menlo Park, CA, USA: AAAI Press. https://www.ijcai.org/proceedings/2018/0095.pdf

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

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

[51] Zhao, X., Pang, Y., Zhang, L., Lu, H., & Zhang, L. (2020, August). Suppress and balance: A simple gated network for salient object detection. In *European conference on computer vision* (pp. 35-51). Cham: Springer International Publishing. [CrossRef]

[52] Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O. R., & Jagersand, M. (2020). U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern recognition, 106*, 107404. [CrossRef]

[53] Pang, Y., Zhao, X., Zhang, L., & Lu, H. (2020, June). Multi-Scale Interactive Network for Salient Object Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 9410-9419). IEEE. [CrossRef]

[54] Wei, J., Wang, S., & Huang, Q. (2020, April). F3Net: 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]

[55] Wei, J., Wang, S., Wu, Z., Su, C., Huang, Q., & Tian, Q. (2020). Label decoupling framework for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13025-13034). [CrossRef]

Farhan Ali is a graduate student currently pursuing a Master's in Computer Science with a specialization in Data Science at Technische Universität Graz, Austria. He has a strong academic background and hands-on experience in data science, machine learning, and computer vision. He worked as an Associate Software Engineer at OpusAI, where he was involved in building user-centric web applications. In addition, he completed data science internships with Oasis Infobyte and Info AidTech, respectively, gaining valuable experience.

Muhammad Ali is an undergraduate student of Software Engineering at the University of Haripur, Pakistan. His research interests include computer vision, machine learning, and artificial intelligence. He has actively contributed to research projects, particularly in the areas of data collection and data annotation, gaining hands-on experience in building high-quality datasets for vision-based applications.

Zaid Muhammad is a Computer Science student with a strong foundation in Python programming and computational problem-solving, complemented by experience in graphic and visual system design. His work focuses on developing intelligent, efficient, and visually robust digital solutions. His research interests include artificial intelligence, machine learning, and technology-driven system design.