Highlights

FEM-Net:Feature Enhancement and Multi-Level Fusion Network for Salient Object Detection

Bingfeng Li, BoXiang Lv, QingShan Chen, XinXin Duan, Xinwei Li

- Introducing a Feature Enhancement and Multi-Level Fusion Network for Salient Object Detection.
- Introducing a multi-scale pooled self-attention module (MPSA) aimed at capturing global salient features by combining multi-scale max pooling with a multi-head self-attention mechanism.
- Introducing an adaptive channel enhancement Block (ACEB) leverages an attention mechanism to assign higher weights to key channels, thereby improving the model's capability to capture fine-grained salient features along the channel dimension.
- Introducing a multi-Level diffusive synergy block (MDSB) that enhances feature interaction across layers by combining a cross-attention mechanism with a dynamic diffusion refinement process to regulate feature propagation and interaction.
- Introducing a dual-domain fusion attention module (DFAM), which effectively models both global and local contexts by integrating a self-attention mechanism with a local enhancement feature extraction unit to optimize feature fusion.

FEM-Net:Feature Enhancement and Multi-Level Fusion Network for Salient Object Detection

Bingfeng Li[a,b], BoXiang Lv[a,*], QingShan Chen[a], XinXin Duan[a] and Xinwei Li[a,b]

[a]School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo, 454003, China
[b]Henan Key Laboratory of Intelligent Detection and Control of Coal Mine Equipment, Jiaozuo, 454003, China
[c]China Resources Wind Power (Wu Gang) Co., Ltd, Pingdingshan, 467000, China

ARTICLE INFO

Keywords:
Salient Object Detection
Feature Enhancement
Multi-Level Fusion

ABSTRACT

Recent advancements in deep learning have substantially accelerated the development of methods for Salient Object Detection. Nevertheless, a fundamental challenge persists: the efficient differentiation between salient objects and background regions, while simultaneously preserving the integrity of global features and minimizing the degradation of local details. In response to these challenges, we introduce a Feature Enhancement and Multi-level Fusion Network specifically designed for Salient Object Detection.To achieve a more comprehensive representation of global features, we present the Multi-Scale Pooling Self-Attention Module, which effectively captures global contextual information by combining multi-scale max pooling across spatial dimensions with a self-attention mechanism. Additionally, to better preserve local details and mitigate the loss of fine-grained information, we propose the Adaptive Channel Enhancement Block, which employs an adaptive weighting strategy to prioritize salient channels, thereby augmenting the model's capacity to capture intricate local features.Moreover, to facilitate efficient interaction between features at different levels, we introduce the Multi-Level Diffusive Synergy Block, which incorporates a cross-attention mechanism that enables deep features to guide shallow features in focusing on salient regions. Furthermore, the dynamic diffusion refinement mechanism we propose enhances the propagation and interaction of multi-level features.To address the risk of deep features disproportionately guiding shallow features, which may result in the loss of local details, we introduce the Dual-Domain Fusion Attention Module. This module synergistically integrates global self-attention with locally enhanced feature extraction units, thereby achieving a balanced optimization between global context modeling and local detail preservation.Extensive experiments conducted across six challenging datasets substantiate the superiority of our model, demonstrating its ability to outperform state-of-the-art Convolutional Neural Network and Transformer-based methods across a range of evaluation metrics. The results indicate significant improvements in both accuracy and boundary detail for Salient Object Detection.

1. Introduction

Salient Object Detection (SOD) focuses on the accurate identification and segmentation of the most visually prominent objects or regions within an image [1, 2]. Due to its extensive applicability in the field of computer vision, SOD plays a pivotal role in various downstream tasks, including object detection [3], visual tracking [4, 5], image retrieval [6], action recognition [7], and stereo matching [8]. Moreover, specialized variants of SOD—such as light field SOD [9, 10, 11], remote sensing SOD [12, 13, 14], and RGB-T SOD [15, 16]—have also been extensively studied within their respective domains.

With the continuous advancements in convolutional neural networks (CNNs) [21, 22], numerous CNN-based SOD methods have been proposed. By leveraging the robust feature extraction capabilities of backbone networks, along with feature enhancement techniques and multi-scale feature fusion, these methods have achieved significant success in accurately identifying salient regions. Notably, BASNet [18], which integrates ResNet-34 [23] into a U-shaped encoder-decoder architecture [24], demonstrates exceptional feature extraction capabilities, enabling the model to capture fine boundary details while maintaining computational efficiency. However, CNN-based SOD methods face challenges in capturing long-range pixel dependencies due to the localized receptive fields of convolution operations. This limitation, coupled with the lack of global contextual guidance, often results in the misclassification of background

*Corresponding author
2123070200008@home.hpu.edu.cn (B. Lv)

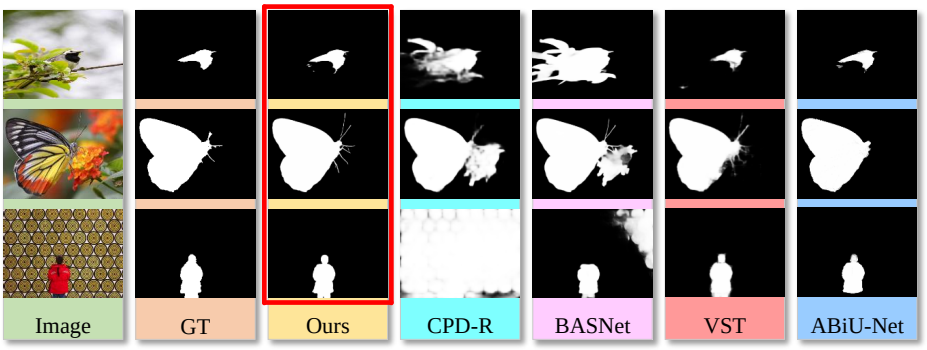

Figure 1: Comparison of various salient object detection methods. Due to the restricted receptive field of convolutional operations, CPD-R [17] and BASNet[18] are prone to erroneously classifying background regions as salient areas. By leveraging global feature dependencies, VST[19] and ABiU-Net[20] effectively reduce the influence of background noise. However, the local details of these methods remain blurred. In comparison, our approach not only accurately distinguishes salient objects from background regions but also captures finer-grained features, even surpassing the quality of the ground truth.

regions as foreground objects. As illustrated in the 4th and 5th columns of Fig. 1, both background regions adjacent to salient areas and those situated farther away are incorrectly identified as salient objects.

Following the substantial success of the Vision Transformer (ViT) [25] in image classification, transformer-based methods for SOD [19, 26] have recently made significant progress. Unlike CNNs, transformers can model long-range feature dependencies through the self-attention mechanism. This allows transformer-based SOD methods to mitigate the influence of background noise on salient object regions. Nevertheless, the sequential processing of input images in ViT [25] results in the inevitable loss of fine-grained features, leading to incomplete detection of salient regions or the omission of local details. As shown in the 6th and 7th columns of Fig. 1, VST [19], which employs T2T-ViT [27] as its backbone, generates incomplete and blurred boundaries for salient regions. ABiU-Net [20], which combines the strengths of both ViT [25] and CNNs, partially addresses the limitations observed in VST [19]; however, significant loss of local detail persists.

To model feature relationships at both local and global scales, M3Net [28] utilizes the Swin Transformer [29] as its backbone architecture and incorporates a multi-scale interaction module to enhance salient regions through cross-scale feature learning. Additionally, by integrating windowed attention and self-attention mechanisms, M3Net [28] effectively preserves fine-grained information. Leveraging the robust feature extraction capabilities of the Swin Transformer[29], M3Net [28] demonstrates outstanding saliency detection performance across several challenging datasets.

In M3Net [28], the shifted window mechanism is employed to capture long-range feature dependencies. While this mechanism facilitates information exchange between adjacent windows, it inherently constrains the model's capacity to capture global contextual information. In addition, M3Net [28] incorporates a hierarchical feature extraction approach, which aids in preserving broader contextual information. However, this approach may result in the smoothing or loss of finer details.

To address the aforementioned problem, we propose a novel Feature Enhancement and Multi-Level Fusion Network (FEM-Net) for SOD. In FEM-Net, we introduce a Multi-Scale Pooled Self-Attention Module (MPSA), which initially performs multi-scale max pooling on the input features to extract abundant salient features. After linearly transforming the input features and concatenated multi-scale salient features into Query, Key, and Value, the self-attention mechanism is employed to capture multi-scale global salient features. However, since MPSA primarily enhances salient features along the spatial dimension, we further augment the model's ability to represent salient features by proposing an Adaptive Channel Enhancement Block (ACEB). ACEB adaptively assigns weights to each channel through an attention mechanism, adjusting these weights based on the importance of the channels. This enhances the model's ability to capture salient features in the channel dimension, enabling it to focus on key features while suppressing irrelevant information.

While the MPSA and ACEB modules effectively extract the salient features of the target, they fail to capture the interactions among features across different layers, which are crucial for a comprehensive representation of salient features. To address this limitation, we propose a Multi-level Diffusive Synergy Block (MDSB). MDSB initially employs a cross-attention mechanism to leverage deep semantic information, guiding shallow features to focus on the salient regions of the semantic target. Subsequently, a dynamic diffusion refinement mechanism regulates the propagation and interaction of features across different levels by introducing a fixed diffusion ratio and an adaptive diffusion matrix, thereby ensuring both the accuracy and flexibility of feature enhancement.

In MDSB, the guidance of shallow features by deep features to focus on high-level semantic information may result in the loss of local details in the shallow features. To address this issue, we propose the Dual-Domain Fusion Attention Module (DFAM). DFAM employs a self-attention mechanism to compute global correlations at each position, enabling the capture of more comprehensive and accurate global contextual information. Simultaneously, it integrates a local enhancement feature extraction unit designed to refine local details. By combining the self-attention mechanism with the local enhancement unit, DFAM achieves a coordinated optimization of both global and local features.

In general, our main contributions can be summarized as follows:

- We introduce a Multi-scale Pooled Self-Attention Module (MPSA), designed to extract global salient features through multi-scale max pooling and multi-head self-attention mechanism. Subsequently, the Adaptive Channel Enhancement Block (ACEB) assigns greater weight to important channels using an attention mechanism, thereby enhancing the model's ability to capture fine-grained salient features in the channel dimension.
- We introduce a Multi-Level Diffusive Synergy Block (MDSB), which facilitates feature interaction across different layers through the synergy of a cross-attention mechanism and a dynamic diffusion refinement mechanism. This mechanism regulates the propagation and interaction of features at various levels.
- To improve prediction accuracy, we design the Dual-Domain Fusion Attention Module (DFAM), which comprehensively models global and local contexts by combining a self-attention mechanism with a local enhancement feature extraction unit, thereby optimizing feature fusion.

2. Related Work

2.1. Salient Object Detection

The origins of salient object detection research can be traced to Liu et al. [30], who developed a saliency detection method grounded in color contrast and spatial information. Achanta et al. [31] made another notable contribution by introducing a saliency detection approach that leverages frequency-tuned low-level visual features, including color. Subsequent advancements integrated background priors [32, 33] and contrast-based hypotheses [34]. However, these methods often depend on handcrafted features [35, 36], constraining both their effectiveness and generalizability. In pursuit of a comprehensive and precise representation, deep learning-based salient object detection methods have progressively emerged as the dominant approach. These methods can be broadly classified into two categories: convolutional neural network (CNN)-based and Transformer-based approaches.

2.2. Convolution Based Methods

In recent years, convolutional neural networks have established themselves as a potent tool for salient object detection, owing to their ability to provide more effective representations of salient features. A prevalent strategy for CNN-based saliency object detection (SOD) involves generating saliency maps through multi-scale feature fusion [37, 38]. Another approach utilizes multi-level contextual guidance to effectively integrate low-level prior cues with high-level features [39]. For instance, Hou et al. [40] introduced skip connections in multi-scale side outputs to integrate shallow and deep features, thereby improving the accuracy of salient feature capture. Zhang et al. [41] proposed a bidirectional information transfer mechanism to optimize the interaction and exchange of multi-scale features extracted by fully convolutional networks. Liu et al. [42] proposed a two-stage saliency object detection (SOD) network, wherein the first stage generates a preliminary saliency map, which is subsequently refined by integrating both global and local contextual information. Similarly, Wang et al. [43] initially employ global features for coarse saliency prediction, which is subsequently refined with local features. Luo et al. [44] integrated multi-scale non-local features to effectively address salient objects of varying sizes, thereby improving the overall performance of salient object detection. Although the CNN-based methods discussed above have demonstrated progress in capturing global contextual information, they

predominantly rely on local features. Furthermore, during the feature extraction and fusion process, the balance between global and local features presents an ongoing challenge.

2.3. Transformer Based Methods

The Transformer, initially developed for natural language processing (NLP) [45], has proven successful in various dense prediction tasks, particularly in salient object detection. For instance, VST [19], which adopts T2T-ViT [27] as its backbone network, incorporates an efficient multi-task decoder to process serialized features, thereby enhancing the effectiveness of visual saliency detection. SRformer [26] integrates the pyramid vision Transformer (ViT) [25] as its encoder and leverages pixel reordering to reconstruct high-resolution features, thereby capturing more distinctive salient details. Nonetheless, their significant reliance on global information may result in the omission of local details, thereby compromising the precision in representing local features. To address these issues, M3Net [28] utilizes Swin Transformer [29] as its backbone network. In contrast to Vision Transformer, the Swin Transformer incorporates a shifting window mechanism that dynamically adjusts the positions of the windows, thereby facilitating the exchange of information between adjacent windows and enhancing the capture of salient features. This design provides M3Net with robust support for feature representation, enabling more accurate localization of salient regions. However, the adjacent windows mechanism constrains the scope of information exchange within M3Net. Furthermore, the downsampling operations in Swin Transformer result in the compression or averaging of fine-grained details, thereby causing the blurring of local features.

2.4. Multi-Scale Feature Aggregation

The aggregation of multi-scale features is essential for generating high-resolution saliency maps. For instance, EGNet [46] incorporates an edge-guided mechanism with multi-scale feature aggregation, thereby enabling the model to capture both low-level edge details and high-level semantic representations. MINet [47] incorporates an interactive learning strategy that optimizes multi-layer feature fusion, thereby enhancing spatial consistency and mitigating feature discrepancies. DGRL [43] employs a recursive learning approach to iteratively refine saliency maps by integrating global contextual information with local details. In contrast, BBRF [48] utilizes a decoder architecture with adjustable receptive fields, thereby facilitating flexible adaptation to objects of varying scales. Nevertheless, these methods predominantly rely on the straightforward aggregation of multi-scale feature maps, thereby failing to fully leverage the intricate relationships among features across scales.

3. Methodology

3.1. Framework overview

The overall architecture of FEM-Net is illustrated in Fig. 2, where the Swin Transformer is employed as the backbone for feature extraction. The feature maps across multiple scales are reconstructed through the MPSA module, thereby refining the global contextual information. The refined features are subsequently processed by the ACEB, which leverages channel attention mechanisms to further enhance the local details of salient features. Subsequently, the MDSB is employed to fuse multi-level features, thereby guiding shallow features to concentrate on the salient regions of semantic targets. Finally, the DFAM optimizes both global context and local features by computing global correlations and enhancing local details, thereby improving the overall feature representation.

3.2. Multi-scale Pooled Self-Attention Module

As illustrated in Fig. 3, given the input features $X \in \mathbb{R}^{N\times C}$, where N represents the token length and C denotes the feature dimension. After a linear transformation with the weight matrix $W_q \in \mathbb{R}^{C\times C}$, the query vector $Q \in \mathbb{R}^{N\times C}$ is derived as follows:

$$Q = XW_q \tag{1}$$

Simultaneously, after reshaping X, we obtain $X \in \mathbb{R}^{C\times H\times W}$, where H and W represent the height and width of the feature map. A max pooling operation is performed along the spatial dimensions with kernel sizes of 3×3, 4×4, and 6×6, thereby generating multi-scale feature maps $V_i \in \mathbb{R}^{C\times H_i\times W_i}$, where i corresponds to the index of the max pooling operation; the V_i are then input into a depthwise convolution for position encoding, followed by an element-wise summation between the position encoding and V_i, we obtain the multi-scale encoded features $V_i^{\text{enc}} \in \mathbb{R}^{N_i\times C}$, where N_i denotes the length of V_i^{enc}. Subsequently, these feature maps are concatenated to obtain a multi-scale token

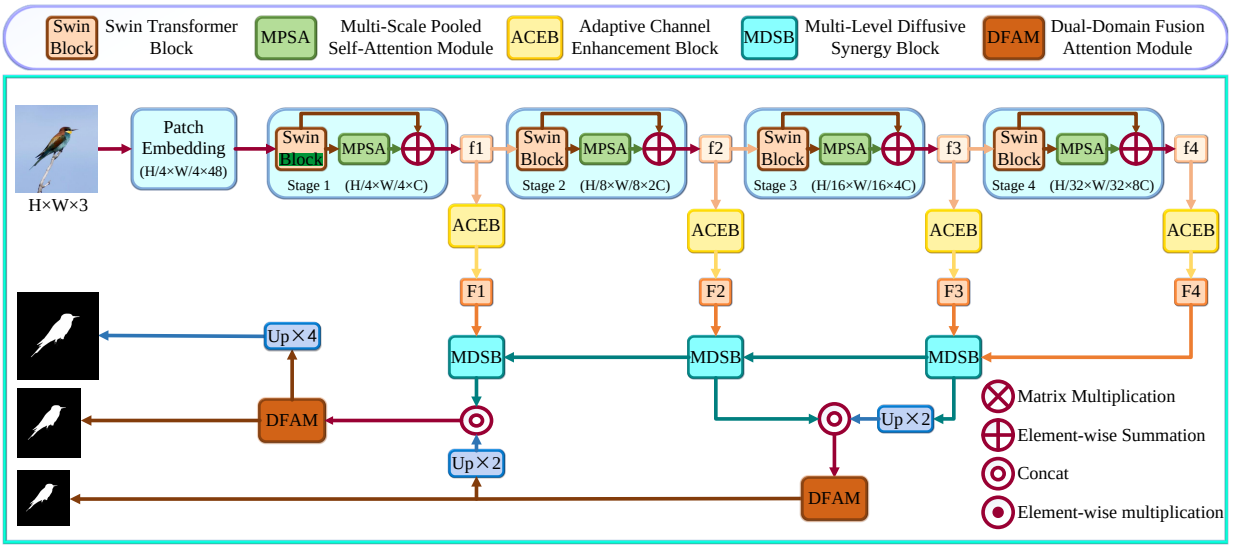

Figure 2: The overall framework of the proposed **FEM-Net** model, which consists of a backbone network and four feature extraction modules. Colored boxes of the same color represent identical feature processing modules. **"Up"** represents the upsampling operations.

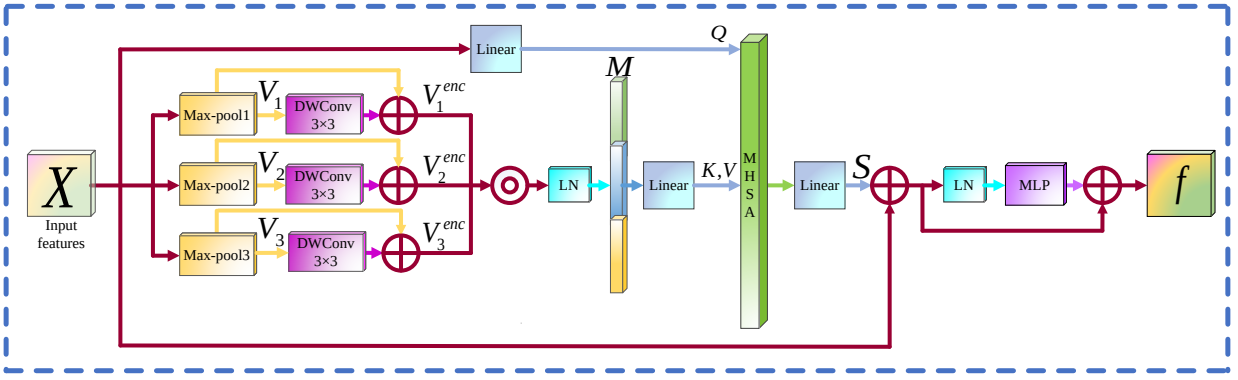

Figure 3: The overall structure of **MPSA**, illustrating the step-by-step process of capturing global features.

feature $M \in \mathbb{R}^{D\times C}$, where $D = N_1 + N_2 + N_3$. A linear transformation is applied to M with the weight matrices $W_k \in \mathbb{R}^{C\times C}$ and $W_v \in \mathbb{R}^{C\times C}$, resulting in the key vector $K \in \mathbb{R}^{D\times C}$ and the value vector $V \in \mathbb{R}^{D\times C}$. This process can be formally expressed as follows:

$$\begin{aligned} V_i &= \text{Maxpool}_i(X) \\ V_i^{enc} &= \text{DWConv}(V_i) + V_i \\ M &= \text{Concat}(V_i^{enc}) \\ K &= MW_k \\ V &= MW_v \end{aligned} \tag{2}$$

With self-attention mechanisms, a new feature representation can be describe as:

$$S = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}}) \cdot V \tag{3}$$

Following the pixel-wise addition of the input features X and S, a residual block incorporating layer normalization and a multilayer perceptron (MLP) is employed to generate the output feature $f \in \mathbb{R}^{N\times C}$, which can be mathematically represented as follows:

$$f = \mathrm{MLP}(\mathrm{LN}(S + X)) + (S + X) \tag{4}$$

In the MPSA module, the multi-scale max pooling operation effectively preserves spatial information across different scales, while the self-attention mechanism facilitates the exchange of information across non-adjacent windows. Consequently, MPSA not only retains the salient features of the target but also substantially enhances the model's capacity to capture global contextual information.

3.3. Adaptive Channel Enhancement Block

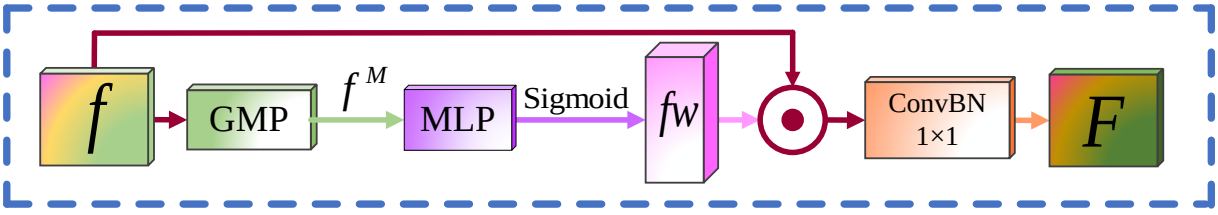

Figure 4: The overall structure of **ACEB**, illustrating the step-by-step process of extracting detailed features.

While MPSA performs exceptionally well in capturing global contextual information, its enhancement remains limited to the spatial dimension. Inspired by CBAM [49], we propose an Adaptive Channel Enhancement Block (ACEB) to more effectively capture the salient features of targets in the channel dimension. As illustrated in Fig. 4, given the feature map $f \in \mathbb{R}^{H\times W\times C}$, the adaptive weight $fw \in \mathbb{R}^{C}$ is obtained by sequentially applying the global max pooling operation, the multilayer perceptron, and the Sigmoid activation function. Then, The adaptive weights fw and f are combined with pixel-wise multiplication. Finally, with the convolution operation and batch normalization, the channel refined saliency map $F \in \mathbb{R}^{H\times W\times C}$ is obtained. The formula presented here summarizes the above process:

$$\begin{aligned} f^{M} &= \mathrm{GMP}(f) \\ fw &= \mathrm{Sigmoid}(\mathrm{MLP}(f^{M})) \\ F &= \mathrm{ConvBN}(fw \odot f) \end{aligned} \tag{5}$$

In summary, ACEB utilizes max pooling to extract distinctive features from each channel, followed by the application of adaptive weighting, thereby enhancing the model's capacity to capture fine-grained details.

3.4. Multi-Level Diffusive Synergy Block

Although the MPSA and ACEB modules effectively capture global contextual information and local details, the lack of mutual guidance between feature layers prevents the optimal representation of multi-scale features. To address this, we propose the Multi-Level Diffusive Synergy Block (MDSB). As depicted in Fig. 5, assume $F_i \in \mathbb{R}^{N_i\times C_i}$ and $F_{i+1} \in \mathbb{R}^{N_{i+1}\times C_{i+1}}$ denote the feature maps, where $N_i = 4N_{i+1}$ and $C_{i+1} = 2C_i$. After the linear transformation and the cross-attention mechanism, the feature vector $G_i \in \mathbb{R}^{N_i\times C_i}$ is obtained. Feeding G_i into the Diffusion layer (DL), which comprises two linear layers and a ReLU activation function, followed by element-wise addition of the diffusion layer's output to G_i, results in the diffused feature vector $G_i^D \in \mathbb{R}^{N_i\times C_i}$.

Simultaneously, the Diffusive matrix $A \in \mathbb{R}^{C\times C}$ is derived by repeating the result of global max pooling (GMP) applied to G_i, followed by multiplication with the Diffusion rate (Dr). The diffused feature G_i^D is then subjected to batched matrix multiplication (BMM) with the Diffusive matrix $A \in \mathbb{R}^{C\times C}$, yielding the updated diffused feature

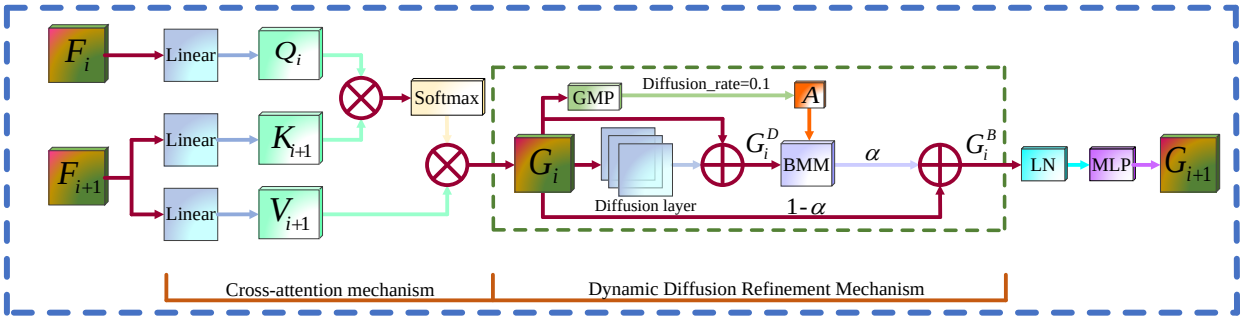

Figure 5: The overall structure of **MDSB** consists of two parts: the cross-attention mechanism and the dynamic diffusion refinement mechanism, which demonstrate the propagation and interaction of features across layers.

$G_i^B \in \mathbb{R}^{N\times C}$. Finally, G_i^B is subjected to a linear transformation, followed by Layer Normalization (LN) and a multi-layer perceptron (MLP), yielding $G_{i+1} \in \mathbb{R}^{N\times C}$. The process can be mathematically represented as follows:

$$\begin{aligned} G_i^D &= G_i + \mathrm{DL}(G_i) \\ A &= \mathrm{repeat}\left(\mathrm{GMP}(G_i) \times \mathrm{Dr}\right) \\ G_i^B &= \alpha \cdot \mathrm{BMM}(G_i^D, A) + (1-\alpha) G_i \\ G_{i+1} &= \mathrm{MLP}(\mathrm{LN}(G_i^B)) \end{aligned} \tag{6}$$

The layer-wise learning mechanism utilized by MDSB enables deep semantic features to guide shallow features in focusing on salient object regions. Especially, with the nonlinear transformations, the dynamic diffusion refinement mechanism captures complex feature relationships, thereby enhancing the richness of feature representations.

3.5. Dual-Domain Fusion Attention Module

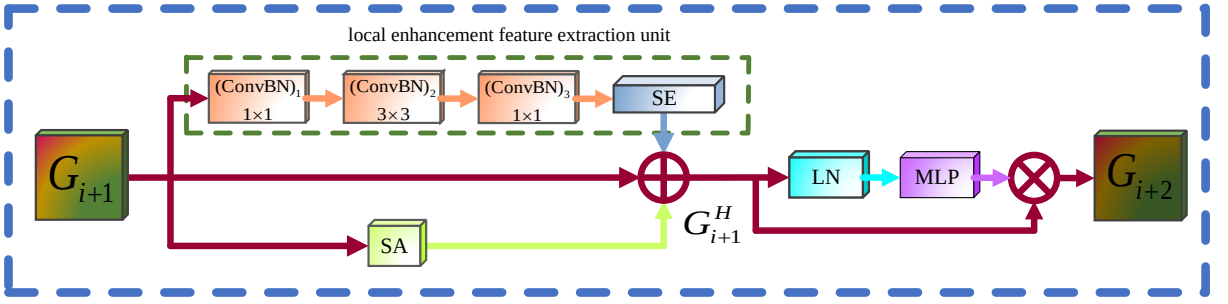

Figure 6: The overall structure of **DFAM**, which aims to model context at both global and local levels.

Although MDSB achieves an optimized expression of multi-scale features through the propagation and interaction of multi-level features, excessive reliance on deep features for guided learning may cause the local details in shallow features to be overshadowed by high-level semantic information. To overcome this, we propose the Dual-Domain Fusion Attention Module (DFAM). As shown in Fig. 6, given the feature map $G_{i+1} \in \mathbb{R}^{N\times C}$, global features and local features are extracted by Self-Attention (SA) and the local enhancement feature extraction unit separately. The local enhancement feature extraction unit consists of three convolution and batch normalization (ConvBN) modules, along with a channel attention module [50]. The global features, local features, and input feature map G_{i+1} are subsequently combined through element-wise addition, yielding fused salient features, denoted as $G_{i+1}^H \in \mathbb{R}^{N\times C}$. The final high-precision saliency map $G_{i+2} \in \mathbb{R}^{N\times C}$ is derived by applying Layer Normalization (LN) and MLP to reweight G_{i+1}^H.

Table 1
The comparison of our FEM-Net with other state-of-the-art SOD across six benchmark datasets. The symbols ↓ and ↑ indicate that lower/higher scores represent better performance, respectively. Red, green, and blue denote the best, second-best, and third-best performances. The suffixes '-R2,''-R,''-V,' and '-S' denote models that utilize Res2Net50[53], ResNet50, VGG16, and Swin Transformer as their backbone networks, respectively.

Dataset	DUT-O			DUTS-TE			ECSSD			HKU-IS			PASCAL-S			SOD		
Method	$M\downarrow$	$S_m\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	$S_m\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	$S_m\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	$S_m\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	$S_m\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	$S_m\uparrow$	$F_\beta^w\uparrow$
CSF-R2[54]	0.055	0.838	0.733	0.037	0.890	0.823	0.033	0.930	0.910	0.030	0.921	0.891	0.069	0.862	0.807	0.098	0.800	0.757
AFNet[55]	0.057	0.826	0.717	0.046	0.855	0.785	0.042	0.914	0.887	0.036	0.905	0.869	0.070	0.849	0.798	0.111	0.774	0.723
ICON-R[56]	0.057	0.844	0.761	0.037	0.889	0.837	0.032	0.929	0.918	0.029	0.920	0.902	0.064	0.861	0.818	0.084	0.824	0.794
POOLNet[57]	0.056	0.836	0.729	0.040	0.871	0.807	0.039	0.921	0.896	0.033	0.917	0.881	0.075	0.832	0.798	0.102	0.797	0.759
U2-Net[58]	0.054	0.847	0.757	0.044	0.861	0.804	0.033	0.928	0.910	0.031	0.916	0.890	0.074	0.844	0.797	0.108	0.786	0.748
DGRL-R[43]	0.063	0.810	0.697	0.051	0.836	0.760	0.042	0.906	0.883	0.037	0.897	0.865	0.074	0.839	0.787	0.106	0.773	0.731
MLMS-V[59]	0.064	0.809	0.681	0.048	0.851	0.761	0.045	0.911	0.871	0.039	0.907	0.859	0.074	0.844	0.779	0.108	0.786	0.726
RAS-V[60]	0.062	0.814	0.695	0.059	0.828	0.740	0.056	0.893	0.857	0.045	0.887	0.843	0.101	0.799	0.736	0.124	0.764	0.720
SRM-R[61]	0.069	0.798	0.658	0.058	0.824	0.722	0.054	0.895	0.853	0.046	0.887	0.835	0.084	0.834	0.758	0.128	0.741	0.670
PFSNet[62]	0.055	0.842	0.756	0.036	0.892	0.842	0.031	0.930	0.920	0.026	0.924	0.910	0.063	0.860	0.819	0.089	0.810	0.781
EGNet[46]	0.053	0.841	0.738	0.039	0.887	0.816	0.037	0.925	0.903	0.031	0.918	0.887	0.074	0.852	0.795	0.097	0.807	0.767
VST[19]	0.058	0.852	0.758	0.037	0.897	0.831	0.032	0.934	0.911	0.029	0.928	0.897	0.060	0.874	0.821	0.085	0.820	0.776
GateNet-R[63]	0.055	0.838	0.729	0.040	0.885	0.809	0.040	0.920	0.894	0.033	0.915	0.880	0.067	0.858	0.797	0.098	0.801	0.753
PWHCNet[64]	0.055	0.850	0.771	0.035	0.898	0.824	0.031	0.932	0.885	0.026	0.929	0.911	0.062	0.866	0.765	—	—	—
ITSD-R[65]	0.061	0.840	0.750	0.041	0.885	0.824	0.034	0.925	0.910	0.031	0.917	0.894	0.066	0.859	0.812	0.093	0.809	0.777
BASNet[18]	0.056	0.836	0.751	0.048	0.866	0.803	0.037	0.916	0.904	0.032	0.909	0.889	0.076	0.838	0.793	0.112	0.772	0.728
LDF[66]	0.052	0.839	0.752	0.034	0.892	0.845	0.034	0.924	0.915	0.028	0.919	0.904	0.060	0.863	0.822	0.093	0.800	0.765
PiCANet[67]	0.054	0.826	0.743	0.040	0.863	0.812	0.035	0.916	0.908	0.031	0.905	0.890	0.064	0.846	0.811	0.094	0.780	0.741
MINet-R[47]	0.056	0.832	0.756	0.039	0.875	0.803	0.032	0.920	0.908	0.028	0.912	0.883	0.065	0.856	0.805	0.102	0.798	0.773
M3Net-R[28]	0.061	0.848	0.769	0.036	0.897	0.849	0.029	0.931	0.919	0.026	0.929	0.913	0.060	0.868	0.827	0.084	0.865	0.819
DSS-V[68]	0.063	0.790	0.697	0.056	0.812	0.755	0.052	0.882	0.872	0.040	0.878	0.867	0.093	0.798	0.759	0.124	0.743	0.710
TCGNet[69]	0.046	0.856	0.789	0.031	0.901	0.856	0.028	0.937	0.936	0.025	0.927	0.919	0.056	0.872	0.845	—	—	—
CPD-R[17]	0.056	0.825	0.719	0.043	0.869	0.795	0.037	0.918	0.898	0.034	0.905	0.875	0.071	0.848	0.794	0.110	0.771	0.713
SRformer[26]	**0.043**	0.860	0.784	0.027	0.910	0.872	0.028	0.932	0.922	0.025	0.928	0.912	0.051	0.878	0.845	0.088	0.809	0.770
BBRF[48]	0.044	0.861	0.803	0.025	0.909	0.886	0.022	0.939	0.944	0.020	0.932	0.932	0.049	0.878	0.856	0.078	0.822	0.802
M3Net-S[28]	0.045	0.872	0.811	0.024	0.927	0.902	0.021	0.948	0.947	0.019	0.943	0.937	**0.047**	0.889	0.864	**0.073**	**0.838**	0.819
Ours	**0.043**	**0.875**	**0.860**	**0.021**	**0.931**	**0.935**	**0.019**	**0.951**	**0.971**	**0.018**	**0.944**	**0.960**	0.058	**0.896**	**0.876**	0.077	0.833	**0.891**

This process can be formally described as follows:

$$\begin{aligned} G_{i+1}^{H} &= G_{i+1} + \mathrm{SA}(G_{i+1}) + \mathrm{SE}(\sum_{i=1}^{3} \mathrm{ConvBN_i}(G_{i+1})) \\ G_{i+2} &= G_{i+1}^{H} \cdot \mathrm{MLP}(\mathrm{LN}(G_{i+1}^{H})) \end{aligned} \tag{7}$$

By combining the self-attention mechanism with the local enhancement feature extraction unit, DFAM achieves a synergistic optimization of both global and local information, providing a more comprehensive, stable, and efficient feature fusion. This facilitates the generation of more precise and detailed saliency maps.

4. Experiments

4.1. Implementation Details and Setup

1) Implementation Details: in our experiments, all input images are firstly resized to 384×384. Random rotation, normalization, and cropping are then applied to augment the dataset. The batch size is set to 3. To effectively capture the complex relationships among features, the RT2T[51, 52] is employed as the upsampling method, where the patch size, stride and padding size are set to [3, 3, 7], [2, 2, 4] and [1, 1, 2], respectively. During the training phase, the Binary Cross-Entropy loss and Intersection over Union loss functions are employed to guide the learning process. The Adam optimizer is used to train the network in two stages. In the first stage, training was conducted for 100 epochs with a learning rate of $1e-4$, and in the second stage, training was carried out for 20 epochs with the learning rate adjusted to $2e-5$.

2)Datasets: The performance of our FEM-Net is assessed across six widely utilized benchmark datasets for saliency object detection, including DUT-OMRON[70], DUTS[71], ECSSD[72], HKU-IS[73], PASCAL-S[74] and SOD[75].

3)Evaluation Metrics: Following previous works, the performance of our FEM-Net is evaluated with the following three distinct evaluation metrics, including:

- **MAE** (Mean Absolute Error) evaluates the average pixel-wise absolute difference between the predicted value P and the ground-truth G, and is defined as:

$$\mathrm{MAE} = \frac{1}{H \times W} \sum_{x=1}^{H} \sum_{y=1}^{W} |P(x, y) - G(x, y)| \tag{8}$$

- **S-measure**(Structure-measure)[76]is employed to assess the structural similarity between predicted results and ground truth. It is computed as:

$$S = (1 - \alpha)S_r + \alpha S_o \tag{9}$$

S_o and S_r denote the object similarity and region similarity, respectively, with α set to 0.5.

- **weighted F-measure**(F_β^w)[77]is an extension of the traditional F-measure, achieved by applying weights w to the various basic metrics. It is defined as:

$$F_\beta^W = \frac{(1 + \beta^2) \cdot \mathrm{Precision}^w \cdot \mathrm{Recall}^w}{\beta^2 \cdot \mathrm{Precision}^w + \mathrm{Recall}^w} \tag{10}$$

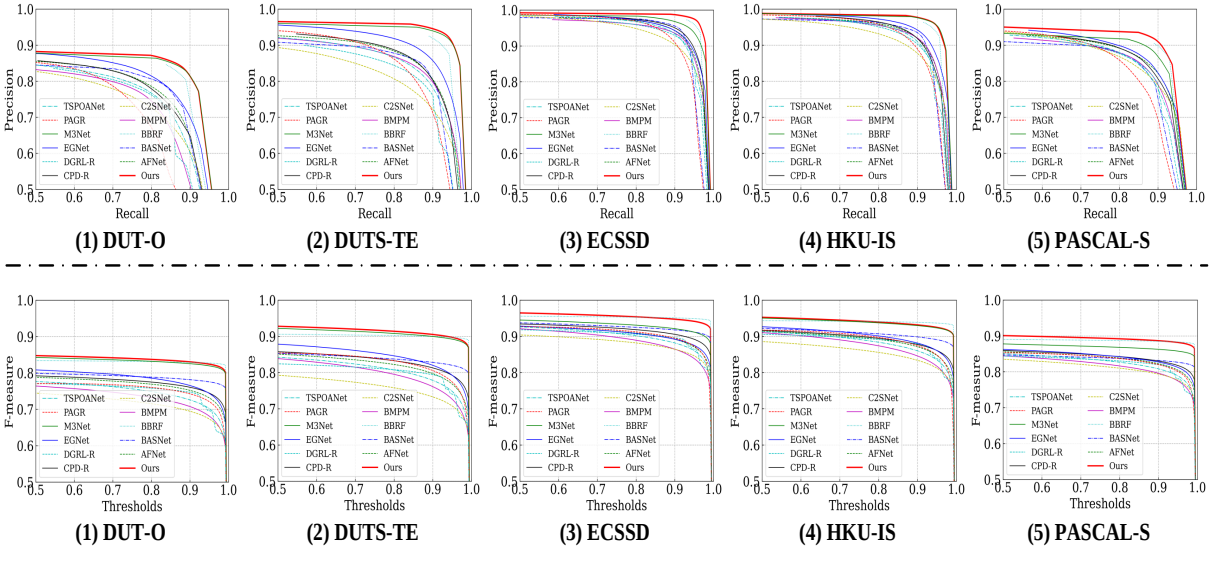

Figure 7: Precision-Recall curves (row 1) and F-measure curves (row 2) of our FEM-Net and other state-of-the-art SOD methods on five benchmark datasets.

4.2. Comparison With State-of-the-Art Methods

1) Quantitative Evaluation: We compare our FEM-Net with other state-of-the-art SOD methods, and report the results in Table 1. To ensure a fair comparison, saliency maps of different methods were generated by rerunning the published code with default parameters. As indicated in Table 1, our FEM-Net outperforms other methods in the majority of cases. This demonstrates the effectiveness of our method. However, in the PASCAL-S and SOD datasets, some performance of our FEM-Net decreased. This is likely due to the limited size of the two datasets, with PASCAL-S comprising 850 images and SOD only 300 images. Consequently, overfitting becomes unavoidable.

To comprehensively evaluate the performance of FEM-Net, we present the PR curve and F-measure curve in Fig. 7, from which it can be observed that our method demonstrates consistently commendable performance.

2)Qualitative Comparison: To intuitively illustrate the performance of FEM-Net, a visual comparison with other methods is presented Fig. 8. The results indicate that FEM-Net consistently produces accurate saliency maps across

a variety of scenarios. Specifically, in rows 1 and 2, the saliency maps generated by FEM-Net exhibit more detailed information, whereas those generated by alternative methods demonstrate lower accuracy and more blurred details, such as the chair's legs and backrest or the grasshopper's legs. In rows 3 and 4, FEM-Net demonstrates its effectiveness in segmenting salient objects within complex backgrounds, whereas other models struggle to suppress background noise. In rows 5, 6, and 7, FEM-Net effectively captures the fine structure of low-contrast salient objects, demonstrating a close alignment with the ground truth. Due to the similarity between the foreground and background, other models either fail to capture parts of the objects or generate maps with low accuracy. The images in rows 8 and 9 contain multiple salient objects, our method accurately segments all salient objects. In contrast, other models only detect a subset of objects, resulting in relatively rough and blurry salient object maps.

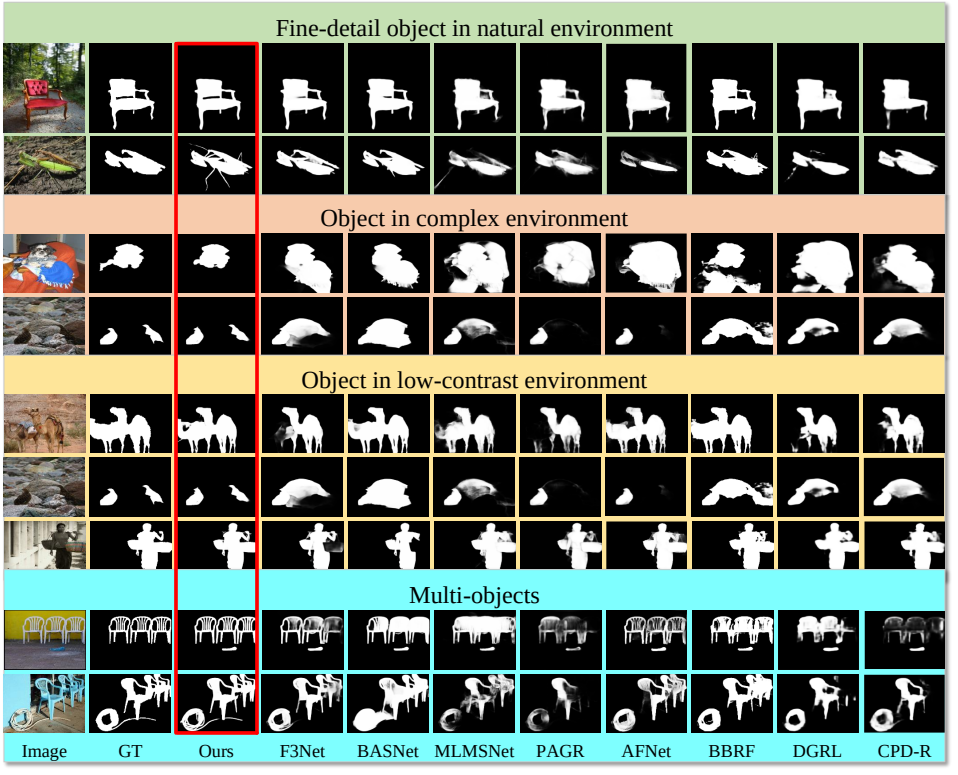

Figure 8: The visual comparison of our FEM-Net model with eight state-of-the-art methods across various scenarios reveals that FEM-Net generates more accurate prediction maps, exhibiting improved boundary and texture details.

4.3. Ablation Study

To demonstrate the effectiveness of different modules in our FEM-Net, we conduct quantitative analysis using several simplified versions of our method. The experimental results on four datasets (DUT-O, DUTS-TE, ECSSD, and HKU-IS) are reported in Table 2, and a visual comparison of each module is provided in Fig. 9.

1) Effectiveness of MPSA: In order to facilitate the acquisition of more salient features, the MPSA was incorporated to augment the model's capacity for global context extraction, as evidenced by the quantitative improvements shown in the "+MPSA" configuration in Table 2 and the corresponding qualitative comparative analysis presented in Fig. 9. To further investigate the efficacy of various pooling strategies in MPSA, we conducted additional experiments to explore the impact of different pooling methods. The results, presented in Table 3 and Fig. 10, demonstrate that max pooling operations exhibit superior capability in preserving salient features, whereas average and mixed pooling are less efficient in this regard.

2)Effectiveness of ACEB: To enhance the representation of salient information from multi-scale features, the ACEB module was utilized to capture finer local details. Its efficacy is evidenced by the "+ACEB" and "+MPSA+ACEB"

Table 2
Ablation study of our FEM-Net. Baseline is Swin Transformer.

ID	Component Settings	DUT-O			DUTS-TE			ECSSD			HKU-IS		
		$M\downarrow$	$S_m\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	$S_m\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	$S_m\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	$S_m\uparrow$	$F_\beta^w\uparrow$
1	Baseline	0.048	0.868	0.823	0.025	0.924	0.906	0.023	0.946	0.942	0.026	0.933	0.931
2	+MPSA	0.047	0.870	0.836	0.024	0.924	0.912	0.023	0.947	0.951	0.024	0.938	0.943
3	+ACEB	0.046	0.871	0.845	0.024	0.925	0.915	0.022	0.947	0.955	0.023	0.939	0.947
4	+MPSA+ACEB	0.046	0.873	0.850	0.023	0.927	0.921	0.021	0.948	0.962	0.021	0.942	0.952
5	+MPSA+ACEB+MDSB	0.045	0.874	0.854	0.022	0.928	0.926	0.020	0.950	0.966	0.019	0.942	0.955
6	+MPSA+ACEB+DFAM	0.044	0.874	0.857	0.021	0.929	0.931	0.020	0.951	0.969	0.019	0.943	0.958
7	**+MPSA+ACEB+MDSB+DFAM**	**0.043**	**0.875**	**0.860**	**0.021**	**0.931**	**0.935**	**0.019**	**0.951**	**0.971**	**0.018**	**0.944**	**0.960**

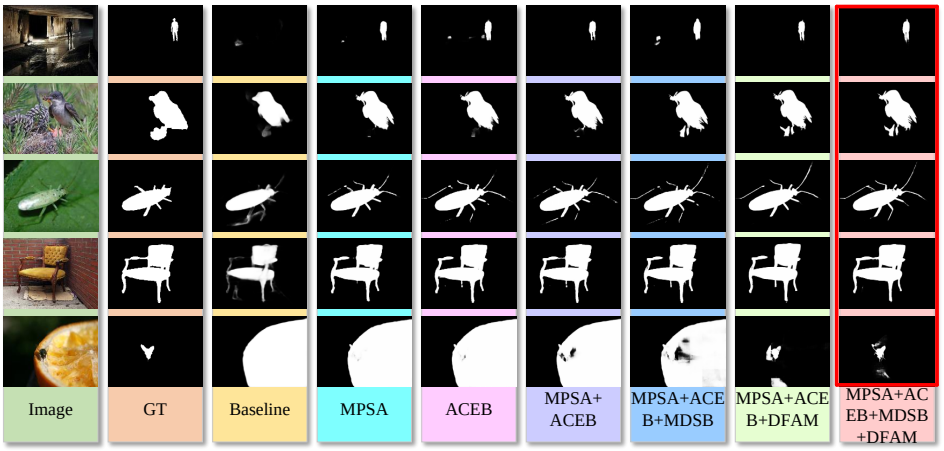

Figure 9: Visual comparison of the ablation study for our proposed FEM-Net.

Table 3
An ablation-based comparative analysis of various pooling methods in the MPSA.

ID	Pooling Methods	DUT-O			DUTS-TE			ECSSD			HKU-IS		
		$M\downarrow$	$S_m\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	$S_m\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	$S_m\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	$S_m\uparrow$	$F_\beta^w\uparrow$
8	Average Pooling	0.044	0.872	0.856	0.024	0.928	0.930	0.020	0.951	0.969	0.019	0.941	0.960
9	Mixed Pooling	0.043	0.874	0.858	0.022	0.929	0.932	0.019	0.950	0.969	0.019	0.943	0.959
10	**Max Pooling**	**0.043**	**0.875**	**0.860**	**0.021**	**0.931**	**0.935**	**0.019**	**0.951**	**0.971**	**0.018**	**0.944**	**0.960**

results presented in Table 2, as well as the visual comparisons of MPSA shown Fig. 9. These results substantiate the effectiveness of the ACEB module.

3) Effectiveness of MDSB: To facilitate deeper interactions between features at different scales and enhance the representation of multi-scale features, MDSB was integrated with MPSA and ACEB. This integration aims to strengthen salient regions while suppressing the influence of non-salient information. The effectiveness of this approach is evidenced by the "+MPSA+ACEB+MDSB" results reported in Table 2, as well as the visual comparisons of MPSA illustrated in Fig. 9. To further analyze the relationship between the effectiveness of MDSB and the number of diffusion layers, a series of experiments were conducted with diffusion layers ranging from 1 to 5. The results presented in Table 4 reveal that MDSB attains optimal performance in enhancing salient regions when configured with 3 diffusion layers.

However, as the number of layers increases beyond this point, performance gradually declines, which may be due to information distortion caused by excessive information diffusion and issues related to gradient vanishing.

4)Effectiveness of DFAM: To enhance the integration of salient information from multi-scale features, DFAM was incorporated on top of MPSA and ACEB to model context at both global and local levels, thereby facilitating

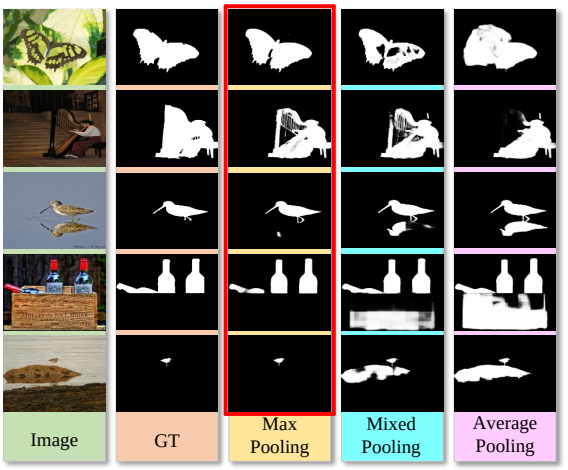

Figure 10: Visual comparison of various pooling methods in the MPSA.

Table 4
Comparative analysis of various diffusion layer configurations in the MDSB.

ID	Diffusion layer	DUT-O			DUTS-TE			ECSSD			HKU-IS		
		$M \downarrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_m \uparrow$	$F_\beta^w \uparrow$
11	1	0.046	0.871	0.857	0.024	0.928	0.932	0.020	0.947	0.966	0.019	0.942	0.955
12	2	0.044	0.873	0.859	0.022	0.931	0.934	0.019	0.950	0.969	0.018	0.943	0.957
13	**3**	**0.043**	**0.875**	**0.860**	**0.021**	**0.931**	**0.935**	**0.019**	**0.951**	**0.971**	**0.018**	**0.944**	**0.960**
14	4	0.045	0.875	0.859	0.023	0.928	0.933	0.021	0.951	0.968	0.019	0.943	0.956
15	5	0.047	0.872	0.858	0.023	0.927	0.931	0.023	0.948	0.967	0.020	0.941	0.954

the generation of high-quality saliency maps. The "+MPSA+ACEB+DFAM" results in Table 2, along with the visual comparisons of MPSA in Fig. 9, confirm the effectiveness of the DFAM module.

5)Effectiveness of Upsampling method and Loss Function: Four widely used upsampling methods—Nearest Neighbor, Bilinear, Sub-pixel Convolution, and Deconvolution—were compared with the fold-with-overlap upsampling method. The results, as shown in Table 5 and Fig. 11, indicate that the fold-with-overlap upsampling method outperforms the other approaches in both evaluation metrics and the visual quality of saliency maps, which exhibit a closer resemblance to the ground truth.

The evaluation of various loss functions employed in the training of FEM-Net is presented in Table 6. Models trained using either BCE loss or IoU loss in isolation did not yield optimal performance. We propose that BCE loss, when applied independently, fails to account for the spatial relationships and contextual dependencies between pixels, thereby limiting its ability to capture the intricate interdependencies among them. In contrast, IoU loss emphasizes the global region, mitigating some of the limitations inherent in BCE loss and enhancing model performance. However, IoU loss is reliant on the overall region overlap, which disregards the smoothness of object boundaries and shape consistency, potentially leading to the loss of local details or imprecise delineation of object edges. To address these shortcomings, a combination of BCE loss and IoU loss was employed, enabling the model to exploit the advantages of both functions, thereby compensating for their individual limitations and resulting in a more robust overall performance.

5. Conclusion

In this paper, we introduce a Feature Enhancement and Multi-Level Fusion Network (FEM-Net) for salient object detection. First, we design a module that combines multi-scale max pooling with self-attention mechanisms

Table 5
Comparative analysis of different upsampling methods.

ID	Upsampling Methods	DUT-O			DUTS-TE			ECSSD			HKU-IS		
		$M \downarrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_m \uparrow$	$F_\beta^w \uparrow$
16	Nearest Neighbor	0.051	0.862	0.842	0.033	0.921	0.925	0.031	0.936	0.952	0.029	0.933	0.946
17	Bilinear	0.046	0.869	0.851	0.026	0.927	0.930	0.023	0.943	0.959	0.024	0.939	0.952
18	Sub-pixel Convolution	0.043	0.870	0.853	0.024	0.929	0.930	0.021	0.946	0.961	0.022	0.942	0.960
19	Deconvolution	0.044	0.872	0.858	0.023	0.929	0.932	0.019	0.948	0.967	0.021	0.944	0.958
20	**Fold with overlap**	**0.043**	**0.875**	**0.860**	**0.021**	**0.931**	**0.935**	**0.019**	**0.951**	**0.971**	**0.018**	**0.944**	**0.960**

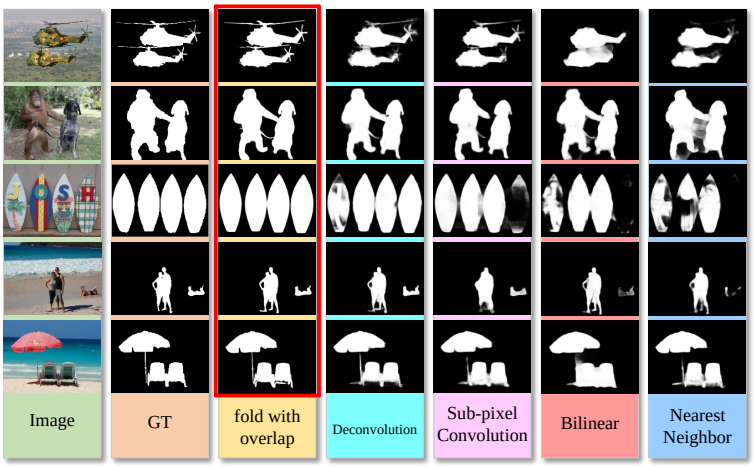

Figure 11: Visual comparison of various upsampling methods. Deconvolution and Sub-pixel Convolution lead to lower prediction quality, while Bilinear and Nearest-Neighbor interpolation cause a noticeable loss of edge and texture details.

Table 6
Performance comparison across different loss settings on DUT-O, DUTS-TE, ECSSD, and HKU-IS datasets.

ID	Loss Setting	DUT-O			DUTS-TE			ECSSD			HKU-IS		
		$M \downarrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_m \uparrow$	$F_\beta^w \uparrow$
21	$\mathcal{L}_{bce}$	0.048	0.869	0.847	0.024	0.927	0.930	0.024	0.942	0.960	0.022	0.937	0.951
22	$\mathcal{L}_{iou}$	0.046	0.873	0.845	0.023	0.926	0.933	0.022	0.947	0.965	0.021	0.940	0.955
23	$\mathcal{L}_{bce} + \mathcal{L}_{iou}$	**0.043**	**0.875**	**0.860**	**0.021**	**0.931**	**0.935**	**0.019**	**0.951**	**0.971**	**0.018**	**0.944**	**0.960**

(MPSA) to extract global salient features along the spatial dimension. Next, we introduce the Adaptive Channel Enhancement Block (ACEB), which assigns higher weights to key channels, thereby enhancing the model's ability to capture local details in the channel dimension. Furthermore, we develop the Multi-Level Diffusive Synergy Block (MDSB), which combines cross-attention mechanisms with a dynamic diffusion refinement process to facilitate the efficient interaction and propagation of features across multiple layers, thereby improving the accuracy and flexibility of salient feature representation. Finally, we propose the Dual-Domain Fusion Attention Module (DFAM), which integrates self-attention with channel-enhanced feature extraction units to optimize the coordination of global and local contextual features, thus improving prediction accuracy. Each of the proposed modules demonstrates significant performance improvements. Extensive experiments across six datasets show that our model outperforms 26 state-of-the-art methods across various evaluation metrics, underscoring its considerable potential for real-world salient object detection applications.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] G. Li, Z. Liu, and H. Ling. Icnet: Information conversion network for rgb-d based salient object detection. *IEEE Transactions on Image Processing*, 29:4873–4884, Mar. 2020.

[2] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang. Review of visual saliency detection with comprehensive information. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(10):2941–2959, Oct. 2019.

[3] Z. Ren, S. Gao, L.-T. Chia, and I. W.-H. Tsang. Region-based saliency detection and its application in object recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(5):769–779, 2014.

[4] Vijay Mahadevan and Nuno Vasconcelos. Saliency-based discriminant tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[5] F. Lin, C. Fu, Y. He, F. Guo, and Q. Tang. Learning temporary block-based bidirectional incongruity-aware correlation filters for efficient uav object tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(6):2160–2174, 2021.

[6] Yue Gao, Meng Wang, Dacheng Tao, Rongrong Ji, and Qionghai Dai. 3-d object retrieval and recognition with hypergraph analysis. *IEEE Transactions on Image Processing (TIP)*, 21(9):4290–4303, 2012.

[7] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 568–576, 2014.

[8] G.-Y. Nie, M.-M. Cheng, Y. Liu, Z. Liang, D.-P. Fan, Y. Liu, and Y. Wang. Multi-level context ultra-aggregation for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3283–3291, 2019.

[9] Y. Chen, G. Li, P. An, Z. Liu, X. Huang, and Q. Wu. Light field salient object detection with sparse views via complementary and discriminative interaction network. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2023.

[10] D. Jing, S. Zhang, R. Cong, and Y. Lin. Occlusion-aware bi-directional guided network for light field salient object detection. In *Proceedings of the ACM Multimedia (ACM MM)*, pages 1692–1701, 2021.

[11] M. Feng, K. Liu, L. Zhang, H. Yu, Y. Wang, and A. Mian. Learning from pixel-level noisy label: A new perspective for light field saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1756–1766, 2022.

[12] C. Li, R. Cong, J. Hou, S. Zhang, Y. Qian, and S. Kwong. Nested network with two-stream pyramid for salient object detection in optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(11):9156–9166, 2019.

[13] Q. Zhang et al. Dense attention fluid network for salient object detection in optical remote sensing images. *IEEE Transactions on Image Processing*, 30:1305–1317, 2021.

[14] Z. Huang, T. Xiang, H. Chen, and H. Dai. Scribble-based boundary-aware network for weakly supervised salient object detection in remote sensing images. *arXiv*, abs/2202.03501, 2022.

[15] Q. Zhang, T. Xiao, N. Huang, D. Zhang, and J. Han. Revisiting feature fusion for rgb-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(5):1804–1818, 2021.

[16] G. Liao, W. Gao, G. Li, J. Wang, and S. Kwong. Cross-collaborative fusion-encoder network for robust rgb-thermal salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7646–7661, 2022.

[17] Z. Wu, L. Su, and Q. Huang. Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[18] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7479–7489, 2019.

[19] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han. Visual saliency transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4722–4732, October 2021.

[20] Y. Qiu, Y. Liu, L. Zhang, H. Lu, and J. Xu. Boosting salient object detection with transformer-based asymmetric bilateral u-net. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(1):1–10, Aug. 2023.

[21] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng. Dual path networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4467–4475, 2017.

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[23] N. Liu, J. Han, and M. H. Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3089–3098, 2018.

[24] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III*, pages 234–241. Springer, 2015.

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[26] Y. K. Yun and W. Lin. Selfreformer: Self-refined network with transformer for salient object detection. 2022.
[27] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and S. Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 558–567, October 2021.
[28] Y. Yuan, P. Gao, and X. Tan. M3net: Multilevel, mixed and multistage attention network for salient object detection. *Journal of LaTeX Class Files*, 14(8):1–12, 2021.
[29] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
[30] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.
[31] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
[32] Yichen Wei, Fang Wen, Wangjiang Zhu, and Jian Sun. Geodesic saliency using background priors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012.
[33] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. Saliency optimization from robust background detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
[34] Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 37(3):569–582, 2015.
[35] Runmin Cong, Jianjun Lei, Huazhu Fu, Qingming Huang, Xiaochun Cao, and Chunping Hou. Co-saliency detection for rgbd images based on multi-constraint feature matching and cross label propagation. *IEEE Transactions on Image Processing (TIP)*, 27(2):568–579, 2018.
[36] Wenguan Wang, Jianbing Shen, Ling Shao, and Fatih Porikli. Correspondence driven saliency transfer. *IEEE Transactions on Image Processing (TIP)*, 25(11):5025–5034, 2016.
[37] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5455–5463, 2015.
[38] Lijun Wang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Deep networks for saliency detection via local estimation and global search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3183–3192, 2015.
[39] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Saliency detection by multi-context deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1265–1274, 2015.
[40] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip Torr. Deeply supervised salient object detection with short. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
[41] Lu Zhang, Ju Dai, Huchuan Lu, You He, and Gang Wang. A bi-directional message passing model for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1741–1750, 2018.
[42] Nian Liu and Junwei Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
[43] Tiantian Wang, Lihe Zhang, Shuo Wang, Huchuan Lu, Gang Yang, Xiang Ruan, and Ali Borji. Detect globally, refine locally: A novel approach to saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3127–3135, 2018.
[44] Zhiming Luo, Akshaya Kumar Mishra, Andrew Achkar, Justin A Eichel, Shaozi Li, and Pierre-Marc Jodoin. Nonlocal deep features for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6000–6010, Dec 2017.
[46] Jian-Xiong Zhao, Ji-Jie Liu, Deng-Ping Fan, Yuhao Cao, Jing Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2019.
[47] Ying Pang, Xiaoyu Zhao, Lei Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
[48] Ming Ma, Chengzhe Xia, Changqing Xie, Xiaowei Chen, and Jia Li. Boosting broader receptive fields for salient object detection. *IEEE Transactions on Image Processing*, 32:1026–1038, 2023.
[49] S. Woo, J. Park, J. Y. Lee, and et al. Cbam: Convolutional block attention module. In *Lecture Notes in Computer Science*. Springer, Cham, 2018.
[50] Jiefeng Hu, Li Shen, Gang Sun, et al. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2017.
[51] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han. Visual saliency transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4722–4732, October 2021.
[52] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and S. Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 558–567, October 2021.
[53] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr. Res2net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(2):652–667, 2021.
[54] S.-H. Gao, Y.-Q. Tan, M.-M. Cheng, C. Lu, Y. Chen, and S. Yan. Highly efficient salient object detection with 100k parameters. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
[55] M. Feng, H. Lu, and E. Ding. Attentive feedback network for boundary-aware salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7107–7116, June 2019.
[56] M. Zhuge, D.-P. Fan, N. Liu, D. Zhang, D. Xu, and L. Shao. Salient object detection via integrity learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
[57] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang. A simple pooling-based design for real-time salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[58] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–12, 2020.

[59] Runmin Wu, Mengyang Feng, Wenlong Guan, Dong Wang, Huchuan Lu, and Errui Ding. A mutual learning method for salient object detection with intertwined multi-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8150–8159, 2019.

[60] S. Chen, X. Tan, B. Wang, and X. Hu. Reverse attention for salient object detection. In *European Conference on Computer Vision (ECCV)*, 2018.

[61] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu. A stagewise refinement model for detecting salient objects in images. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4039–4048, 2017.

[62] M. Ma, C. Xia, and J. Li. Pyramidal feature shrinking for salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1–9, 2021.

[63] X. Zhao, Y. Pang, L. Zhang, H. Lu, and L. Zhang. Suppress and balance: A simple gated network for salient object detection. In *European Conference on Computer Vision (ECCV)*, 2020.

[64] Q. Zhang, M. Duanmu, Y. Luo, Y. Liu, and J. Han. Engaging part-whole hierarchies and contrast cues for salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(6):3644–3658, Jun. 2022.

[65] H. Zhou, X. Xie, J.-H. Lai, Z. Chen, and L. Yang. Interactive two-stream decoder for accurate and fast saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[66] J. Wei, S. Wang, Z. Wu, C. Su, Q. Huang, and Q. Tian. Label decoupling framework for salient object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[67] N. Liu, J. Han, and M.-H. Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[68] Xiaoyong Hu Ali Borji Zhenwei Tu Hou, Mengmeng Cheng and Philip H. S. Torr. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[69] Yi Liu, Ling Zhou, Gengshen Wu, Shoukun Xu, and Jungong Han. Tcgnet: Type-correlation guidance for salient object detection. *IEEE Transactions on Intelligent Transportation Systems*, 2023. accepted for publication.

[70] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3166–3173, 2013.

[71] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan. Learning to detect salient objects with image-level supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[72] Q. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[73] G. Li and Y. Yu. Visual saliency based on multiscale deep features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5455–5463, 2015.

[74] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille. The secrets of salient object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 280–287, 2014.

[75] V. Movahedi and J. H. Elder. Design and perceptual validation of performance measures for salient object segmentation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops (CVPRW)*, pages 49–56, 2010.

[76] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4548–4557, 2017.

[77] R. Margolin, L. Zelnik-Manor, and A. Tal. How to evaluate foreground maps. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2014.