



Attention enhanced machine instinctive vision with human-inspired saliency detection

Habib Khan ^a, Muhammad Talha Usman ^a, Imad Rida ^b, JaKeoung Koo ^{a,*}

^a School of Computing, Gachon University, 1342 Seongnam-daero, Seongnam-si 13120, South Korea

^b Laboratoire Biomécanique et Bioingénierie UMR 7338, Centre de Recherches de Royallieu, Université de Technologie de Compiègne, Compiègne, France

ARTICLE INFO

Dataset link: <https://github.com/habib1402/MFRNet-SOD>

Keywords:

Saliency detection
Visual intelligence
Multi-scale features
Attention
Progressive fusion

ABSTRACT

Salient object detection (SOD) enables machines to recognize and accurately segment visually prominent regions in images. Despite recent advancements, existing approaches often lack progressive fusion of low and high-level features, effective multi-scale feature handling, and precise boundary detection. Moreover, the robustness of these models under varied lighting conditions remains a concern. To overcome these challenges, we present Attention Enhanced Machine Instinctive Vision framework for SOD. The proposed framework leverages the strategy of Multi-stage Feature Refinement with Optimal Attentions-Driven Framework (MFRNet). The multi-level features are extracted from six stages of the EfficientNet-B7 backbone. This provides effective feature fusions of low and high-level details across various scales at the later stage of the framework. We introduce the Spatial-optimized Feature Attention (SOFA) module, which refines spatial features from three initial-stage feature maps. The extracted multi-scale features from the backbone are passed from the convolution feature transformation and spatial attention mechanisms to refine the low-level information. The SOFA module concatenates and upsamples these refined features, producing a comprehensive spatial representation of various levels. Moreover, the proposed Context-Aware Channel Refinement (CACR) module integrates dilated convolutions with optimized dilation rates followed by channel attention to capture multi-scale contextual information from the mature three layers. Furthermore, our progressive feature fusion strategy combines high-level semantic information and low-level spatial details through multiple residual connections, ensuring robust feature representation and effective gradient backpropagation. To enhance robustness, we train our network with augmented data featuring low and high brightness adjustments, improving its ability to handle diverse lighting conditions. Extensive experiments on four benchmark datasets — ECSSD, HKU-IS, DUTS, and PASCAL-S — validate the proposed framework's effectiveness, demonstrating superior performance compared to existing SOTA methods in the domain. Code, qualitative results, and trained weights will be available at the link: <https://github.com/habib1402/MFRNet-SOD>.

1. Introduction

Saliency detection (SD) aims to replicate the natural ability of the human visual system (HVS) to distinguish the most prominent objects in a visual scene instinctively. Just like human eyes are drawn to prominent features, the same SD systems work on how machines perceive and process visual information. Salient object detection (SOD) was defined by path-finding researchers as a method to mimic the human visual system's ability to quickly identify areas of interest in a scene [1]. This pioneering model simulated the effortless way humans identify regions of interest, bridging the gap between cognitive processes and artificial visual perception. SD marks a pivotal advancement in computer vision (CV), leveraging traditional image processing and advanced deep learning (DL) models. It reduces computational load by focusing on key

areas of interest, making processes more efficient [2]. The benefits of SD extend to various fields, offering improved user experiences, driving innovations in autonomous vehicles by enhancing obstacle detection and navigation, and enriching interactive media by highlighting crucial content.

SD functions as an important stage in CV, optimizing computational resource allocation and improving the performance of diverse image analysis tasks across multiple domains. It can be applied to segmentation as co-saliency [3], object recognition with its foundational established role since early research [4], image captioning [5], retrieval [6], image understanding [7], pixel-level image fusion [8]. Visual tracking [9] by focusing on the most relevant regions. In robotics, SD allows machines to interact more effectively with their environment

* Corresponding author.

E-mail addresses: habibkhan@gachon.ac.kr (H. Khan), jakeoung@gachon.ac.kr (J. Koo).

<https://doi.org/10.1016/j.imavis.2024.105308>

Received 16 July 2024; Received in revised form 22 September 2024; Accepted 17 October 2024

Available online 4 November 2024

0262-8856/© 2024 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

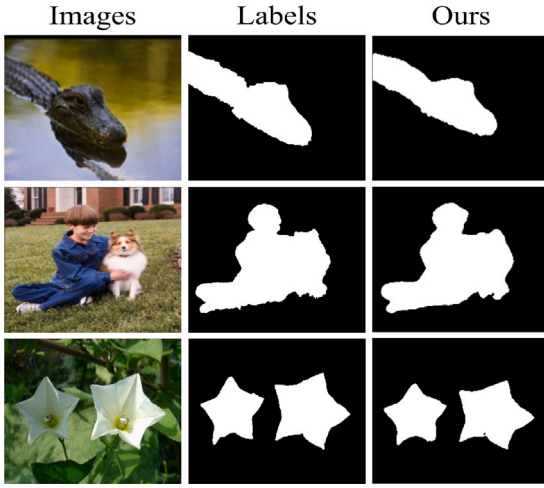


Fig. 1. Results demonstrate the efficacy of our proposed network (Ours) with ground truth and images in complex environments. From top to bottom: (a) Partial SOD in low contrast settings (crocodile in water), (b) Multi-object with overlapping subjects (child and dog), and (c) Flowers in low light with background leaves.

for complex vision navigation [10]. In medical imaging, it aids in identifying critical areas within scans, improving diagnostic accuracy and treatment planning [11]. Additionally, SD algorithms optimize visual perception in autonomous vehicles by prioritizing regions of interest and enhancing object detection and scene understanding [12].

SOD can be categorized into several types: video-based SOD, RGB-D SOD, RGB SOD, and RGB-Thermal (RGB-T) SOD. Video-based SOD focuses on identifying salient regions in dynamic sequences by leveraging both spatial and temporal information. RGB-D SOD enhances accuracy in complex scenes by incorporating depth data to capture spatial relationships better. RGB-T SOD combines visual and thermal images, offering robust performance across diverse lighting conditions. RGB based SD, relying solely on color information, is widely used for real-time applications due to its simplicity and effectiveness, serving as the foundation for more advanced methods. Earlier SOD approaches [1,13–18] often utilized heuristic methodologies and handcrafted low-level feature practices. These methods are less effective in complex environments with low contrast and varying object scales.

In recent years, the extensive applications of convolutional neural networks (CNNs) have significantly advanced the field of SOD. Furthermore, attention mechanisms are thoroughly investigated in CV [19] and are now progressively included in SOD models to improve feature representation and contextual awareness. Initial CNN-based approaches utilized superpixels or region proposals as fundamental units for prediction, assigning a uniform saliency score to all pixels within the same superpixel or proposal [20,21]. Following this, fully convolutional networks (FCNs) were introduced to improve detection performance, enabling more efficient pixel-wise saliency prediction by extracting multi-level features from images [22–24]. Several studies have focused on fusing multi-level features, typically through summation or concatenation, to achieve sharp boundaries in the saliency maps (SMs) [25–28]. Some SOD methods employed single-scale CNN layers followed by max-pooling layers to extract optimal features [29]. Additionally, there has been interest in exploring SOD using dilated CNNs [30–32]. For instance, the method [30] proposed a contextual pyramid module that uses dilated CNNs with different dilation rates to capture global context. It then applies channel compression to the convolutional features before generating the output. The study [31] incorporated dilated layers with a dilation rate of 2 to capture global context. Similarly, [32] explored convolutions with varying dilation rates, while [33] introduced an attention-driven contextual feature fusion model.

1.1. Limitations in the related studies

Despite significant advancements in FCN-based methods [22–24], several challenges persist. A primary limitation is the suboptimal use of multi-scale features, often not refined at appropriate network stages. While deeper layers encode rich semantic information and shallower layers retain finer details, shallow features in some studies [25–28] may be noisy, and the existing fusion attempts may degrade performance. Additionally, repeated upsampling and pooling operations hinder the preservation of sharp object boundaries in predicted SMs. Single-scale convolutions also struggle to manage salient object position and size variability effectively, as noted in [29]. Although dilated CNN-based methods [30–32] show promise, they often struggle to address the local information loss introduced by higher dilation rates. Furthermore, existing methods frequently need more effective progressive feature fusion and refinement strategies. Although [33] introduced attention-over-attention mechanisms, they lack in validating the use of intermediate-level features for progressive network construction, limiting the effective combination of high-level semantics and low-level details. Lastly, the robustness of existing methods under varying lighting conditions remains a concern. Most must adequately tackle the challenges posed by complex environments, which restricts their practical applicability in real-world scenarios.

To address the highlighted limitations, we introduce the Multi-Scale Feature Refinement strategy with a Dual Attention-Driven framework called MFRNet. The proposed framework is specifically designed to tackle the challenges of SD as illustrated in Fig. 1. Our contributions are as follows:

1.2. Contributions

- Instead of relying on single-scale intermediate features, we extracted multiscale feature maps from six stages of the EfficientNet-B7 backbone. This approach offers a comprehensive representation of salient objects, empirically validated by achieving a favorable balance between computational efficiency and performance. By leveraging the advantages of compound scaling, EfficientNet-B7 provides a set of features at different scales, capturing both detailed low-level information and rich high-level semantic cues essential for precise SOD.
- We proposed the Spatial-optimized Feature Attention (SOFA) module to refine three initial-stage feature maps from the EfficientNet-B7 backbone. These features are then passed from the Convolution Feature Transformation (CFT) with 1×1 and 3×3 convolutions, followed by spatial attention mechanisms to refine spatial information. The SOFA module concatenates and upsamples these refined features, producing a comprehensive spatial representation. This design enables the module to generate sharper and more precise SMs, distinguishing salient objects from complex backgrounds.
- The Context-Aware Channel Refinement (CACR) block is designed to collect contextual information at several scales efficiently. This is achieved using dilated convolutions with optimized dilation rates at feasible framework points. This practiced approach ensures a wider field of view without losing important features, enhancing the ability to detect salient objects across different scales. The subsequent channel attention (CA) process refinement highlights informative channels, further boosting the detection performance.
- The proposed framework employs a progressive feature fusion strategy, fusing low and high-level semantic and spatial details through multiple concatenative skip connections. This ensures continuous information flow and preserves original and attention-refined features. This strategic use of residuals and attention mechanisms allows the framework to learn the most important

features, leading to a robust and comprehensive feature representation. Finally, the SOFA features are multiplied element-wise with the output of CACR, enabling an adaptive feature refinement that combines spatial and context-aware channel information.

- We trained the framework using augmented data with low and high brightness adjustments to simulate complex environments and enhance robustness. This approach ensures the model can effectively handle varying lighting conditions, improving its generalization capabilities. Extensive experimental analyses are conducted on ECSSD, HKU-IS, DUTS, and PASCAL-S datasets. The results show the effectiveness of the integrated modules and outperformed state-of-the-art (SOTA) networks in the domain.

The paper is organized in the following parts: Section 2 extensively examines the current body of literature. Section 3 elaborates on the components utilized in the suggested framework. Section 4 shows the validation of the proposed steps followed by the performance evaluation using the benchmarks and the comparative analysis with SOTA methods. Finally, Section 5 includes a conclusion and recommendations for future paths in the SD domain.

2. Related work

Researchers have made significant progress in the challenge of SOD in recent decades. Generally, SOD-related literature can be classified into two main groups, which include conventional SOD methods and deep feature-based SOD approaches. These approaches are briefly discussed in the below two subsections.

2.1. Handcrafted feature-based conventional methods

Traditional SOD approaches heavily depend on intuitive priors and manually designed low-level features [1]. Early methods [17,34] leveraged handcrafted features and often utilize region proposals or superpixels as the basic unit for SOD. Many traditional saliency studies considered visual contrast from global or local perspectives. The saliency of each pixel or superpixel is determined by comparing the variations in center-surround among several low-level features, such as spatial [35], pattern [36], color [17,37], and structure [38]. The work [39] introduced an approach for generating a SM by analyzing local contrast and using a fuzzy growing technique to identify important regions in the map. They introduced a feasible system for analyzing image attention. This framework includes three levels of attention analysis: attended view, attended regions, and attended points. It enables the automated extraction of attention from images, similar to how humans perceive visual information. Another study [40] shows that a merged approach, including face detection and low-level saliency, surpasses a low-level framework in accurately predicting where humans will focus their attention while seeing images of natural scenes. The work showed that observers fixate on faces with high probability, even without specific instructions, and the model's performance is not impaired by spurious face detector responses, suggesting a bottom-up face detection mechanism. However, these methods have limitations in their scope of comparison and ability to extract the edges of important objects. The local contrast-assisted approaches [39,40] can only detect edges with high contrast. These methods often cannot fully capture the salient object due to their restricted focus on small areas. To overcome these limitations, global approaches have been proposed that calculate the saliency of a pixel or region by evaluating its distinctiveness compared to the whole image. The work [41] introduced a computationally efficient approach that exploits luminance and visual color features. The proposed algorithm outperforms five SOTA methods in both ground-truth evaluation and SOD application, achieving higher precision. The work [17] proposed a simple multi-scale algorithm that generates SMs, which are utilized to initialize SaliencyCut for unsupervised salient object segmentation. The proposed work outperforms fifteen networks

on various datasets and enables effective sketch-based image retrieval by extracting salient object masks from Internet images. The authors in [37] introduced a method to consider cues from a scale perspective, measuring region-based scales instead of varying patch sizes. The ultimate saliency scores are determined by optimally integrating saliency information from various scales through a hierarchical inference process, which chooses single-scale features to generate the final SM. Their utilized approach enhances detection quality on traditionally challenging images, and the authors also constructed an extended ECSSD dataset. Some studies [42,43] have combined multiple features, leading to improved detection performance of simple visual looks. However, these methods face challenges when dealing with complex images with cluttered backgrounds. To address this limitation, researchers have explored the incorporation of high-level priors based on human attention mechanisms as additional guidance. Some studies proposed Boundary-prior-based methods [44–46] to effectively suppress background pixels by taking features of both background and foreground cues. For instance, the work [45] introduced an SD algorithm that combines dense and sparse reconstruction techniques using background. The study [47] introduced a method that generates maps by assessing the similarity between image regions and foreground and query templates. Another study by [46] employed a model based on absorbing Markov chain for boundary-prior-based SD. In addition to boundary priors, the objectness prior, which emphasizes regions that exhibit object-like characteristics, has been utilized in several works [48,49]. When these priors are combined with human cognitive experience, the resulting methods often improve performance in cases that align with the priors. However, a potential drawback of relying heavily on priors is that they may not be universally applicable to all scenarios, leading to low performances when the image content deviates from the assumptions made by the priors.

2.2. Deep learning-assisted SOD approaches

CNN-based approaches excel in pattern recognition due to their ability to capture effective feature extraction [50–52]. Building on this, attention mechanisms have also gained significant traction in image and video analysis tasks [53–55]. Recently, CNNs have shown exceptional ability in capturing and expressing complex semantic information, leading to impressive improvements in SOD methods [31, 56]. Deep feature-based saliency methods can be categorized into FCN and patch-based approaches. Earlier approaches [20,21,57,58] utilized deep semantic features that were primarily patch-based. These methods commonly consider each segmented area as a fundamental component for calculating SMs, attributing an identical value to all pixels belonging to the same segment. These patch-based methods typically involve extracting CNN features from each patch, followed by a classification or regression step to assign saliency scores. These techniques, although surpassing handcrafted feature-based methods, were hindered by the loss of spatial details and frequently overlooked fine-grained information present in the early layers of the network. Additionally, the need to run multiple times for each patch made them computationally expensive. FCN-based methods were introduced to predict salient objects by considering pixel-wise features to address these limitations. The approaches used features gained from many levels of the FCN and deployed diverse multi-level integration methodologies to produce SMs at the pixel level. For instance, the study [23] used an attention-based method to consider side-output residuals for top-down effective SOD. The utilized reverse attention involves generating attention maps at each side-output level and propagating them back to the previous level, enabling the network to refine the SMs progressively. The work [59] introduced short connections to skip-layer structures to exploit multi-scale features fully. They employed skip connections, ultimately allowing the network to integrate semantic features from high-level deep layers with low-level details from shallow

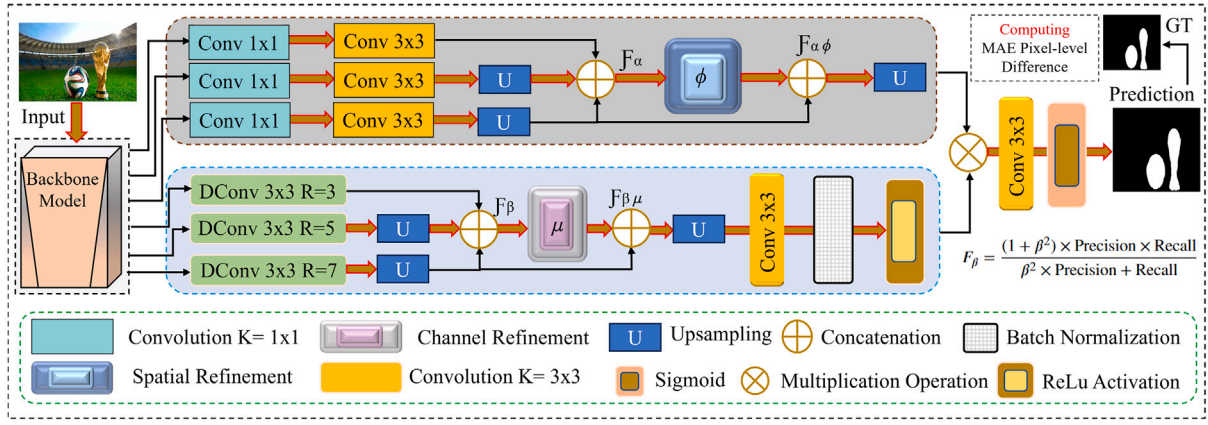


Fig. 2. Visual overflow of the proposed MFRNet.

layers, resulting in more precise SMs. Several works [60–62] incorporated edge labels to assist in generating SMs with clear boundaries. These approaches attempt to enhance the spatial consistency and sharpness of the produced SMs by training the network to predict object boundaries in addition to saliency. The research [63] introduced a module for detecting prominent edges, which aims to precisely identify salient boundaries and improve the segmentation of salient objects while retaining their edges. This module consists of an edge detection branch that operates in parallel with the SD branch, allowing the network to optimize for both tasks jointly. The study [64] enhanced their pooling-based network by including an edge detection branch, which effectively enhances the fine details of prominent objects by simultaneous training. Attention mechanisms have also been employed to focus more on salient regions while exploiting multi-scale saliency information [65]. Feasible integration of attention modules assisted the model in adaptively weighing the importance of different regions and feature channels. Additionally, some studies [31,66,67] have proposed emphasizing boundaries in the loss function. The study [66] used a loss function inspired to punish inconsistencies on prominent edges. This incentivizes the neural network to generate SMs with well-aligned borders. The work [31] proposed a loss function that takes object contours by incorporating structural similarity index, binary cross-entropy, and intersection over union. This hybrid loss function aims to simultaneously optimize the pixel-wise accuracy and the structural coherence of the generated SMs, effectively capturing local and global saliency characteristics. These advancements in CNN-based SOD networks have enhanced the ability to capture and integrate multi-level and multi-scale features. However, challenges remain in effectively balancing the contributions of different feature levels and maintaining clear object boundaries, leaving room for further research and development in this field.

3. Proposed framework

In this section, we introduce the key components of our proposed framework designed for efficient SOD as shown in Fig. 2. We begin with an overview of the technical aspects of the architecture. Subsequently, we delve into the details of deep multiscale feature extraction, highlighting the extraction of features at multiple levels for comprehensive feature representation. Following this, we elaborate on the SOFA module with a deep analysis of CFT and spatial refinement via SA. Additionally, we discuss the CACR module, which leverages dilated convolutions with optimized dilation rates to capture contextual information effectively. Lastly, we explore the multilevel utilization of residuals and progressive feature fusion techniques employed to refine and integrate features across different scales.

3.1. Architecture overview

The Multi-Scale Feature Refinement with Optimal Dual Attention-Driven Framework (MFRNet) aims to improve SOD through the efficient use of multi-scale feature extraction, attention mechanisms, and progressive feature fusion. MFRNet is based on the EfficientNet-B7 backbone and extracts six feature maps from its reduction phases, capturing essential low-level and high-level information for precise SD. The first main component of the framework is SOFA module which processes the initial three feature maps $\mathbb{X}^{F_1} \in \mathbb{R}^{b \times 32 \times 112 \times 112}$, $\mathbb{X}^{F_2} \in \mathbb{R}^{b \times 48 \times 56 \times 56}$, $\mathbb{X}^{F_3} \in \mathbb{R}^{b \times 80 \times 28 \times 28}$ undergo CFT using 1×1 and 3×3 convolutions, followed by upsampling of the $\mathbb{X}^{F_2}$ and $\mathbb{X}^{F_3}$ to match the first stage spatially features $\mathbb{X}^{F_1}$, resulting in a concatenated feature map F_α with the dimensions $\in \mathbb{R}^{b \times 160 \times 112 \times 112}$. The F_α features are passed to the SA block ϕ to produce the spatially refined features and are then concatenated with the F_α to produce $F_{\alpha\phi}$ with dimensions $\in \mathbb{R}^{b \times 320 \times 112 \times 112}$. The features $F_{\alpha\phi}$ are upsampled to retain the initial spatial dimensions 224×224 before the final progressive feature fusion.

The second CACR module processes channel-wise rich feature maps and passes the three mature features to the dilated convolution blocks 3×3 with optimal dilation rates. The features $\mathbb{X}^{F_5} \in \mathbb{R}^{b \times 640 \times 7 \times 7}$ and $\mathbb{X}^{F_6} \in \mathbb{R}^{b \times 2560 \times 7 \times 7}$ are upsampled to align with $\mathbb{X}^{F_4} \in \mathbb{R}^{b \times 224 \times 14 \times 14}$. These features are concatenated to form F_β with the Ch = 3424 and spatial dimension 14×14 . This feature map is passed through the CA module, improving the channel-wise features. The F_β features are concatenated with the CA-attended μ . The resultant feature map $F_{\beta\mu}$ are upsampled to produce the feature map 224×224 same to the version of $F_{\alpha\phi}$. These features are fed to the 3×3 convolution block to reduce the channel size to 320. The output of SOFA and CACR features are then fused element-wise using multiplication operation to generate a robust feature map, which undergoes a final 3×3 convolution. This strategy is adopted to reduce $\in \mathbb{R}^{b \times 1 \times 224 \times 224}$ and feed it to the sigmoid activation to generate the final SM with the effective representation of both spatial and semantic details.

3.2. Deep multiscale feature extraction

Multiscale feature extraction has proven highly effective as it captures diverse detail and contextual information, enhancing model performance. The ability to extract features from multiple scales is crucial in SOD as we need to strengthen the features effectively. We utilized EfficientNetB7 for intermediate multi-level feature extraction, where different layers capture varying levels of detail and abstraction. EfficientNetB7 is a SOTA CNN-based architecture known for balancing accuracy and computational efficiency [68]. It employs a balanced scaling approach, increasing the network's depth, width, and resolution to enhance performance while maintaining computational efficiency. In our proposed framework, we utilize feature maps extracted from six

key reduction stages, referred to as $\mathbb{X}_{1---n}$ where $n = 6$. The network progressively refines features across six reduction stages. Starting from 112×112 spatial dimensions with 32 channels, it gradually reduces to 56×56 (48 channels), 28×28 (80 channels), and 14×14 (224 channels). The final two stages maintain 7×7 dimensions, with 640 channels, then expand to 2560. This structure balances spatial detail and semantic depth, capturing fine-grained and high-level features. It is observable that the extracted features can be categorized into low-level, mid-level, and high-scale features. The low-level features originate from layers one and two and are crucial for recognizing complex information and edges, which are vital for accurately defining the boundaries of prominent objects. The mid-level features are derived from the third layer, preserving complex patterns and textures. In addition, the upper-level features from layers four to six contain rich semantic details and contextual clues, effectively allowing the network to differentiate between prominent objects from the surrounding environment. Our framework integrates features from different sizes to retain complex and broad knowledge, which can be further improved to accurately identify and locate important objects in diverse circumstances and settings.

3.3. Spatial-optimized Feature Attention (SOFA)

The SOFA module is designed to refine the initial spatially rich three extracted feature maps. Initially, the feature maps extracted from the first three layers of the backbone network undergo a CFT process. This involves passing each feature map through a sequence of 1×1 and 3×3 convolutional blocks to enhance their representational power. The 1×1 convolutions efficiently adjust channel depths, enabling cross-channel interactions and feature compression. Meanwhile, the 3×3 convolutions capture local spatial contexts, enhancing edge and texture representations. This combination allows the framework to learn more compact yet informative feature representations, balancing computational efficiency and richness. Notably, features from the second $\mathbb{X}^{F_2}$ and third $\mathbb{X}^{F_3}$ features are upsampled to align with the spatial dimensions of the first layer, facilitating an integration of multi-scale spatial information. Subsequently, the transformed feature maps $\mathbb{X}^{F_{1,2,3}}$ are concatenated to produce a unified representation as shown in Eq. (1).

$$F_\alpha = \text{Conc}(\text{CFT}(\mathbb{X}^{F_1}), \text{Up}(\text{CFT}(\mathbb{X}^{F_2})), \text{Up}(\text{CFT}(\mathbb{X}^{F_3}))) \quad (1)$$

3.3.1. Spatial refinement

After the CFT process, the SA refinement ensures the precise delineation of boundaries between prominent objects and their backgrounds. The output from the CFT (concatenated feature map) then undergoes processing through a spatial refinement module. The concatenated features of the initial three stages are represented as F_α . We practiced the strategy similar to [30,69] for the increase in the receptive field without increasing parameters. We applied two convolution layers with batch normalization for the high-level feature according to the operations shown in Fig. 3. Next, we apply the sigmoid function to the encoded spatial feature map to perform normalization processing.

The spatial refinement selectively enhances the most relevant spatial regions within the while suppressing less informative areas. Importantly, the spatially refined features are integrated into the original concatenated features using a concatenative (Conc) skip connection strategy as shown in Eq. (2).

$$F_{\alpha\phi} = \text{Conc}(F_\alpha, \phi(F_\alpha)) \quad (2)$$

To align with the input resolution, we upsample (Up) $F_{\alpha\phi}$ to 224×224 as shown in Eq. (3), preparing it for subsequent final feature fusion.

$$F_{\text{SOFA}} = \text{Up}(F_{\alpha\phi}, \text{Spatial size} = (224, 224)) \quad (3)$$

The SOFA module facilitates edge refinement by sharpening the boundaries of salient objects, thereby improving the precision of SMs. Moreover, it enhances features, which is crucial for accurately distinguishing salient objects from complex backgrounds.

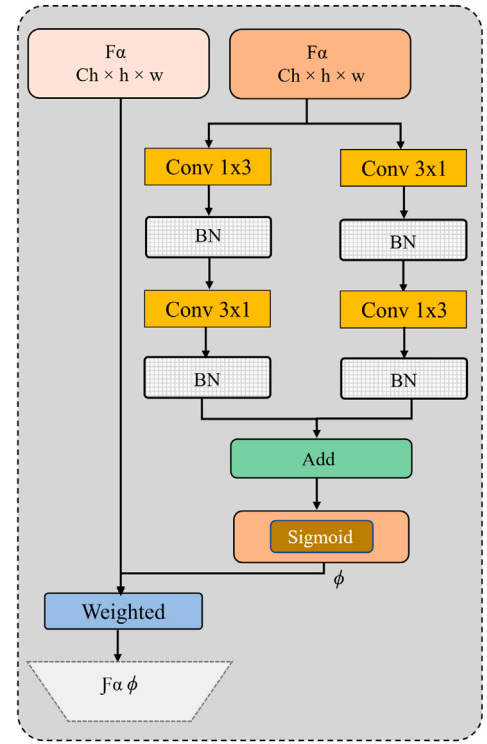


Fig. 3. Spatial refinement block integrated into the SOFA module of the proposed framework.

3.4. Context-Aware Channel Refinement (CACR)

The backbone extracted features $\mathbb{X}^{F_4}$, $\mathbb{X}^{F_5}$, and $\mathbb{X}^{F_6}$ are processed through dilated convolutional layers with rates 3, 5, and 7 respectively. This gathers contextual information from a wider field of view, which is crucial for segmenting salient objects across different scales. $\mathbb{X}^{F_5}$ and $\mathbb{X}^{F_6}$ are upsampled to match the dimensions of $\mathbb{X}^{F_4}$, then concatenated along the channel dimension as shown in Eq. (4).

$$F_\beta = \text{Concat}(\mathbb{X}^{F_4}, \text{Upsample}(\mathbb{X}^{F_5}), \text{Upsample}(\mathbb{X}^{F_6})) \quad (4)$$

3.4.1. Channels refinement

Channel and SA mechanisms are widely used to enhance feature selection by highlighting important channels and spatial regions [70, 71]. In our framework, the concatenated feature map F_β undergoes CA to highlight informative channels and suppress less relevant ones. This refined map is combined with the original features through a concatenative skip connection. This preserves the rich channel information while integrating the refined context. We unfold high-level features F_β as $F_\beta = \mathbf{X}^{F_4}, \mathbf{X}^{F_5}, \dots, \mathbf{X}^{F_5}$, where $\mathbf{X}_i^F \in \mathbb{R}^{Ch \times w \times h}$ is the i th of $\mathbf{X}^F$ and Ch shows the number of channel. Adaptive average pooling (AP) is applied to each feature map, followed by two fully connected layers to capture channel interdependence (Fig. 4). To control complexity and improve generalization, a common approach is to encode the channel-wise feature vector using a bottleneck structure consisting of two fully connected layers surrounding a non-linear activation function, as in [30,69]. This encoding process helps to compress the channel-wise information into a compact representation. Subsequently, a sigmoid function is applied to normalize the encoded channel-wise feature vector, constraining the values within a fixed range. The normalization step ensures that the learned channel-wise weights have a consistent scale and prevents the network from assigning excessive importance to any particular channel. The block produces the final output $F_{\beta\mu}$ by

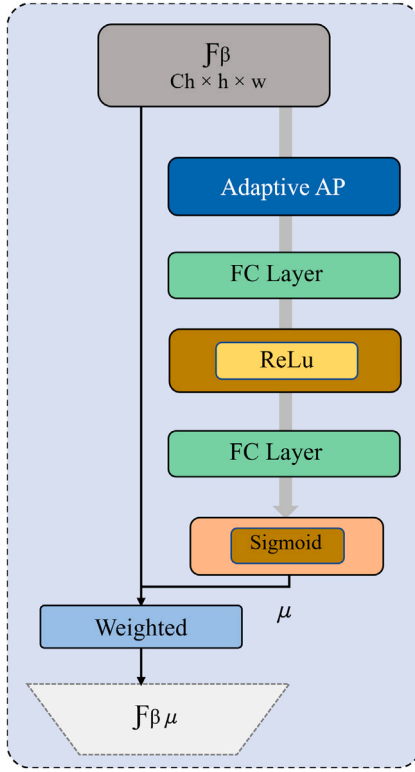


Fig. 4. Visual overflow of the integrated channel refinement module in CACR stage.

combining the features F_β with $CA(F_\beta)$ using a weighting method as shown in Eq. (5).

$$F_{\beta\mu} = \text{Conc}(F_\beta, CA(F_\beta)) \quad (5)$$

Finally, F_{CACR} features are acquired as shown in Eq. (6) and Visually illustrated in Fig.

$$F_{\text{CACR}} = \text{Conv3} \times 3(\text{Up}(F_{\beta\mu}, \text{size} = (224, 224))) \quad (6)$$

3.5. Progressive features fusion

The Progressive Feature Fusion (PFF) strategy in our framework effectively combines high-level semantic information with low-level spatial details by strategically using residual connections and attention mechanisms. This approach preserves the original context while refining features, which is crucial for accurate SOD. Our framework implements residual connections at key stages: between the concatenated features $\mathbb{X}^{F_{1:2:3}}$ and the SA module output; combining SA module output with original concatenated features; connecting features $\mathbb{X}^{F_4}$, $\mathbb{X}^{F_5}$, and $\mathbb{X}^{F_6}$ with channel-refined features. Finally, we fused the output features of the SOFA and CACR modules using element-wise multiplication operation. It is observable that the element-wise multiplication enables adaptive feature refinement, preserving spatial structure while introducing non-linear interactions, leading to improved performance compared to concatenation. The progressive enhancement approach enables our framework to build a robust and comprehensive feature representation by preserving initial context and iteratively refining features. The PFF strategy ensures an effective combination of multi-level features, preservation of original spatial and channel information, and enhanced gradient flow during backpropagation. This refined architecture allows our framework to leverage low-level details and high-level semantics effectively.

4. Experiments

This section offers an extensive overview of the utilized datasets, data augmentation methods, experimental configuration, and evaluation metrics. It elaborates on the in-depth studies and empirical validations of the proposed framework. Finally, it conducts a comprehensive qualitative and quantitative analysis to evaluate the superiority of the proposed framework.

4.1. Datasets

In our study, we trained and evaluated the proposed framework across four widely used benchmark datasets in SOD. These datasets include ECSSD [72], HKU-IS [21], PASCAL-S [49] and DUTS [73]. The utilized datasets present unique properties, such as complex structures, multiple disconnected salient objects, and objects with uncertain environments to challenge the SOD networks. ECSSD includes 1000 images precisely annotated with semantically significant structures, making it appropriate for assessing the ability to recognize fine features in salient objects. HKU-IS contains 4447 challenging images with multiple disconnected objects, overlapping boundaries, and low color contrast, making accurate segmentation difficult. PASCAL-S contains 850 images with distinct salient objects against complex backgrounds, offering a diverse set for rigorous evaluation. We also used the large-scale DUTS benchmark, which contains 15,572 annotated images. Table 1 provides a detailed overview of these datasets, highlighting key properties such as the number of images, training and test splits used in our specific experiment, and detailed characteristics. Moreover, Fig. 5 presents a selection of representative images from these datasets, illustrating the diverse range of complexities and environments. These datasets collectively serve as robust benchmarks, enabling thorough assessment and validation of the proposed framework across different setups.

4.1.1. Enhancing data variation

We applied a random brightness function to enhance and intentionally lower the brightness and contrast of images in a stochastic manner. This practice serves as a critical data augmentation technique to diversify the dataset. Initially, this function adjusts the contrast by scaling each pixel value with a randomly generated factor ranging from 0.5 to 1.5, thereby reducing or enhancing the contrast of the images. Subsequently, it modifies the brightness by adding a random integer within the range of -20 to 19 to each pixel, further altering the visual features of the data. To ensure the integrity of the images, the function clips all pixel values to remain within the valid range of 0 to 255. This randomized augmentation strategy significantly enhances the framework's robustness by exposing it to a diverse array of modified images, potentially improving its performance on unseen datasets. After applying the techniques, the samples are shown in Fig. 6, where images with high and low brightness adjustments are visualized.

4.2. Experimental setup

The experiments were performed on a system equipped with an Intel(R) Core(TM) i9-10900X CPU (10 cores) and 192 GB of RAM. Model training was accelerated by the PyTorch framework using an NVIDIA GeForce RTX 4090 GPU with 24 GB of VRAM, all within a Windows 10 environment. Key libraries such as torch, torchvision, NumPy, Matplotlib, Pandas, and Pillow were used to support the development and training of the framework. Training configurations involved exploring epochs, batch sizes, optimizers, and learning rates. Optimal results were achieved with 50 epochs and an input image size of 224×224 , using a batch size 8 to balance performance and efficiency. Data augmentation techniques such as low and high-intensity brightness were applied to enhance generalization across diverse visual conditions. It is observable that for fair comparisons, we followed the standard way of previous methods to use the training set of DUTS to train and evaluate

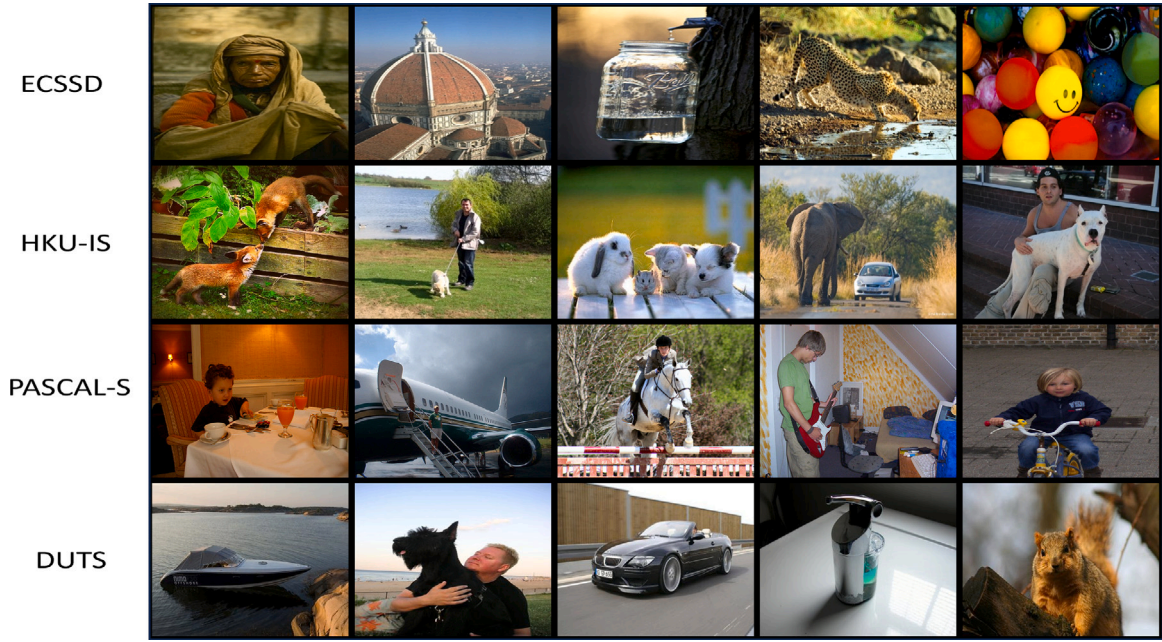


Fig. 5. Example images from the four datasets used in our study. The samples illustrate a range of challenging visual scenarios.

Table 1

Summary of the utilized datasets and train/test split of the data in the study.

Dataset name	Description	Total images	Train split	Test split	MAE	f_{β}
ECSSD	Complex, cluttered scenes	1000	800	200	0.0215	0.9575
HKU-IS	High-resolution, rich structures	4447	4000	447	0.0200	0.9383
PASCAL-S	Real-world, multiple objects	850	700	150	0.0273	0.9041
DUTS	Large-scale, diverse scenes	15,572	10,553	5019	0.0318	0.8754



Fig. 6. Sample images including with low and high brightness adjustments.

ECSSD, PASCAL-S, HKU-IS, and DUTS-TE. Additionally, we conducted experiments within each dataset using the train/test splits outlined in Table 1. Furthermore, the Adam optimizer was employed with Momentum = 0.9, learning rate = 0.0002 (determined empirically), and weight decay = 0.0005, chosen based on experimental insights and

Table 2

Comparative analysis of different backbones including EfficientNet (EffNet) family in terms of standard evaluation metrics of MAE and f_{β} measures.

Backbone	ECSSD		HKU-IS	
	MAE	f_{β}	MAE	f_{β}
MobileNet	0.1192	0.8971	0.0877	0.8863
ShuffleNet	0.1026	0.8995	0.0862	0.8941
MobileNetV2	0.0778	0.9052	0.0673	0.9014
ShuffleNetV2	0.0733	0.9174	0.0545	0.9086
EffNetB0	0.0523	0.9205	0.0458	0.9113
EffNetB1	0.0462	0.9288	0.0441	0.9134
EffNetB2	0.0428	0.9304	0.0408	0.9144
EffNetB3	0.0395	0.9326	0.0381	0.9177
EffNetB4	0.0379	0.9363	0.0343	0.9218
EffNetB5	0.0327	0.9415	0.0305	0.9242
EffNetB6	0.0282	0.9441	0.0279	0.9296
Ours	0.0215	0.9575	0.0200	0.9383

literature in SOD. The optimal selection of these experimental settings and hyperparameters significantly contributed to improve the accuracy of the proposed framework.

4.3. Evaluation metrics

Two well-established evaluation metrics, MAE and F measure f_{β} , are used for assessing performance. These metrics provide comprehensive insights into the effectiveness of the framework across diverse evaluation dimensions. MAE is a widely used evaluation metric in SOD that measures the mean absolute difference between the predicted SM and the ground truth (GT). A lower MAE shows higher performance, as it demonstrates that the predicted saliency scores are closer to the GT. Conversely, a higher MAE defines larger discrepancies between the

Table 3

Ablation study using different combinations of the integrated modules to assess the progressive improvement in the network.

Settings	ECSSD		HKU-IS		PASCAL-S	
	MAE	f_β	MAE	f_β	MAE	f_β
Baseline	0.0321	0.9167	0.0380	0.9000	0.0386	0.8640
Baseline + LL	0.0305	0.9261	0.0367	0.9115	0.0375	0.8655
Baseline + LL + CFT	0.0295	0.9330	0.0362	0.9158	0.0367	0.8671
Baseline + LL + CFT + SA	0.0289	0.9377	0.0354	0.9206	0.0358	0.8673
Baseline + LL + CFT + SA + Residual	0.0288	0.9381	0.0331	0.9224	0.0348	0.8685
Baseline + LL + CFT + SA + Residual + HL	0.0279	0.9428	0.0317	0.9289	0.0342	0.8722
Baseline + LL + CFT + SA + Residual + HL + SC	0.0273	0.9452	0.0298	0.9310	0.0321	0.8750
Baseline + LL + CFT + SA + Residual + HL + DC	0.0265	0.9474	0.0292	0.9344	0.0310	0.8786
Baseline + LL + CFT + SA + Residual + HL + DC + CA	0.0261	0.9476	0.0289	0.9345	0.0297	0.8791
Baseline + LL + CFT + SA + Residual + HL + DC + CA + Residual	0.0241	0.9501	0.0269	0.9373	0.0284	0.8832
Baseline + LL + CFT + SA + Residual + HL + DC + CA + Residual + AUG (Ours)	0.0215	0.9575	0.0200	0.9383	0.0273	0.9041

GT and predicted SMs. In contrast, the f_β strikes a balance between precision and recall. In order to calculate f_β , precision, and recall are established through the comparison of binary maps produced from different thresholds used to the predicted and GT SMs. The f_β incorporates a harmonic mean of recall and precision scores, adopting referred-to standards in the domain, with β^2 being assigned a value of 0.3.

$$f_\beta = \frac{(1 + \beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}} \quad (7)$$

To calculate the MAE, the absolute differences between the pixels in the predicted SM and those in the GT are summed up. This sum is then divided by the total number of pixels in the image to yield the final normalized MAE value. This provides a direct measure of the average magnitude of errors between the predicted and actual saliency values, thereby quantifying the accuracy of the predictions relative to the GT:

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

where Y means the GT SM, and P denotes the predicted SM. These evaluation metrics are rigorously computed using Python toolboxes, ensuring robust and dependable assessment outcomes. By leveraging both MAE and f_β , we achieve a thorough assessment of our model's efficacy in precisely identifying salient objects across various datasets and scenarios.

4.4. Intermediate feature analysis

We conducted an empirical analysis to identify the optimal backbone for intermediate feature extraction. Table 2 presents a comprehensive assessment of various backbone models integrated with our proposed modules. The performance of these models is evaluated using two key metrics: MAE and the f_β measure. Among the evaluated models, MobileNetV2 achieved an MAE of 0.0778 and an f_β score of 0.9052 on ECSSD. ShuffleNetV2 performed slightly better with an MAE of 0.0733 and an f_β of 0.9174. The EfficientNet series demonstrated consistent improvements, with EffNetB0 to EffNetB6 achieving MAEs ranging from 0.0523 to 0.0282 and corresponding f_β scores increasing from 0.9205 to 0.9441. Our proposed MFRNet framework outperformed all evaluated backbones, achieving an MAE of 0.0215 and an f_β score of 0.9575 on ECSSD, and an MAE of 0.0200 with an f_β score of 0.9383 on the HKU-IS dataset. These results underscore the effectiveness of integrating our proposed modules with the optimal backbone architectures and the novel idea of extracting features from the targeted layers. These practices affirm the capability to enhance feature representation and improve SOD performance across various visual scenarios.

4.5. Ablation study

In our comprehensive ablation study, we systematically examined various combinations of integrated modules to evaluate their impact on the performance as tabulated in Table 3. These practices are conducted

across three benchmark datasets: ECSSD, HKU-IS, and PASCAL-S. Starting from the intermediate features extracted from the backbone (baseline), we incrementally added modules to assess improvements in both MAE and F_β measure, which balances precision and recall. Initially, integrating Low-Level Features (LL) modestly enhanced performance across all datasets, laying a foundational improvement in capturing finer details and edges within salient objects. Subsequent integration of the CFT module continued to refine predictions, particularly benefiting from improved spatial localization on intricate images from ECSSD and complex scenes in HKU-IS.

Introducing feature refinement using SA modules marked a significant improvement, which is crucial for accurately detecting disconnected and overlapping salient objects in HKU-IS. Further enhancement through Residual modules underscored improved boundary delineation and overall prediction accuracy. Integrating High-Level Features (HL) expanded the comprehensive understanding of the framework, demonstrating substantial gains across all evaluation metrics. Additionally, incorporating Standard Convolution (SC) and Dilated Convolution (DC) modules improved the ability of the framework to capture contextual information across different scales, further enhancing accuracy. Continued refinement with CA modules highlighted informative feature channels and strengthened the discriminative capabilities of the framework. It is also evident in [74] that spatial and channel attention mechanisms progressively enhance the network's performance by refining feature representation and contextual awareness [54]. Adding residual connections to these changes substantially strengthened feature representation, boosting the ultimate framework's performance and resilience. Finally, the integration of Data Augmentation (AUG) techniques, including adjustments in brightness with high and low-level lights, significantly boosted the ability of the model to generalize across diverse visual scenarios, achieving superior scores in both MAE reduction and F_β measure across all evaluated datasets.

4.6. Qualitative analysis

Visual representations provide valuable insights into the performance of our proposed framework across various scenarios. Fig. 7 compares input images, GT labels, and predictions generated by the proposed framework on utilized benchmark datasets. MFRNet accurately delineates salient objects against complex backgrounds, showcasing its ability to capture complex details and maintain sharp boundaries. Fig. 8 presents a comprehensive visual comparison of our framework's predictions with input images, GT labels, and other SOTA networks. It is evident from these visual comparisons that MFRNet achieves competitive performance, outperforming other methods in accurately highlighting salient objects across diverse datasets. It is observable that our framework leverages a multi-scale feature refinement approach combined with optimal dual attention mechanisms, enhancing its capability to capture both low-level details and high-level semantic information effectively. This strategic integration allows our framework to refine spatial features using the SOFA module and capture multi-scale contextual information through CACR blocks. Progressive feature



Fig. 7. Visual illustrations of comparing input images, labels, and our framework prediction.

Table 4

Comparison of computational complexity (Params) and Frames Per Second (FPS) of the proposed network with various SD methods.

Methods	Params (M)	FPS
RFCN [78]	134.69	0.4
DCL [79]	66.24	1.4
BRN [24]	126.35	3.6
PiCANet [27]	47.22	5.6
DSS [59]	62.23	7.0
Amulet [25]	33.15	9.7
EGNet [60]	111.64	9.9
DHSNet [80]	94.04	10
SRM [22]	43.74	12.3
C2S [81]	137.03	16
NLDF [66]	35.49	18.5
MINet [82]	47	22.3
AFNet [63]	35.95	28.4
CPD [75]	47.85	32.4
BASNet [31]	87.06	36.2
MFRNet (Ours)	74.41	38.28

fusion ensures seamless information integration across network layers, contributing to robust and precise SD results. These visual analyses underscore the effectiveness of our network in handling complex scenes and diverse salient objects, thereby demonstrating its superiority over existing methods in the field of SOD.

4.7. Computational complexity and efficiency

In our computational complexity and efficiency evaluation, we compared our proposed framework against several SOTA methods, as shown in Table 4. Each method was assessed based on two key metrics [85]: the number of parameters (Params in millions) and the FPS achieved during inference. Our approach achieved a parameter count of 74.41 million, positioning it competitively alongside existing methods. Importantly, the model demonstrated robust performance in real-time scenarios, achieving an FPS of 38.28. This efficiency is crucial for applications requiring rapid processing of SD tasks, ensuring timely and responsive outputs without compromising on accuracy. Among the benchmarked methods, MFRNet compares favorably in terms of both parameter efficiency and computational speed. For instance, while some methods like RFCN and DCL offer lower FPS rates due to their higher parameter counts, others like BASNet and CPD prioritize higher FPS at the cost of increased parameter complexity. Our approach balances these factors, leveraging its optimized architecture to deliver competitive performance across varied computational demands. These results underscore the practical viability of our framework for real-world applications, offering a blend of computational efficiency and robust SOD capabilities.

4.8. Comparative analysis

Table 5 summarizes the performance comparison of our proposed MFRNet with several SOTA approaches across four benchmark datasets: ECSSD, HKU-IS, PASCAL-S, and DUTS-TE. MAE and F_β measures are used as standard evaluation metrics. Our framework achieved remarkable results across all datasets and outperformed the SOTA approaches. Particularly, it reaches an MAE of 0.030 and a F_β of 0.936 on ECSSD, demonstrating higher accuracy in comparison to other approaches, including CPD [75] and DFI [83]. The proposed framework demonstrates comparable performance on the HKU-IS, PASCAL-S, and DUTS-TE datasets, with MAE values of 0.023, 0.058, and 0.031, and corresponding F_β scores of 0.929, 0.878, and 0.875, respectively. The proposed framework shows more consistent performance compared to other SOTA methods like BASNet [62] and ACFFNet [33], which achieve MAE/ F_β scores of 0.037/0.927 and 0.032/0.935, respectively, on the ECSSD dataset. Our framework outperforms these benchmark methods, highlighting its ability to detect salient regions with high precision effectively.

5. Conclusion

In this paper, we proposed a novel framework designed for salient object detection (SOD), leveraging the strengths of EfficientNet-B7 for multiscale feature extraction. The proposed framework extracts detailed low-level and rich high-level features from six stages to offer optimal multiscale features. We practiced the successful integration of the Spatial-optimized Feature Attention (SOFA) block to enhance spatial features selectively. Moreover, the Context-Aware Channel Refinement (CACR) block refines multi-scale contextual information effectively. Our framework utilizes a progressive feature fusion technique to merge high-level information with low-level spatial features for strengthening before ultimate prediction. This is achieved by several concatenative skip connections, resulting in a robust and detailed feature representation. To enhance the robustness, we augment the data with varying brightness levels to simulate complex environmental conditions. Our rigorous experiments on widely used benchmark datasets show that the proposed framework consistently outperforms state-of-the-art methods in accuracy and robustness for SOD tasks. In the future, we intend to divert our research focus to optimize the network for real-time applications and extending its capabilities to tackle even more complex environments. Moreover, we plan to propose more challenging large-scale benchmarks. Additionally, we will propose more advanced attention mechanisms to further boost the precision and robustness across a wide range of environmental conditions.

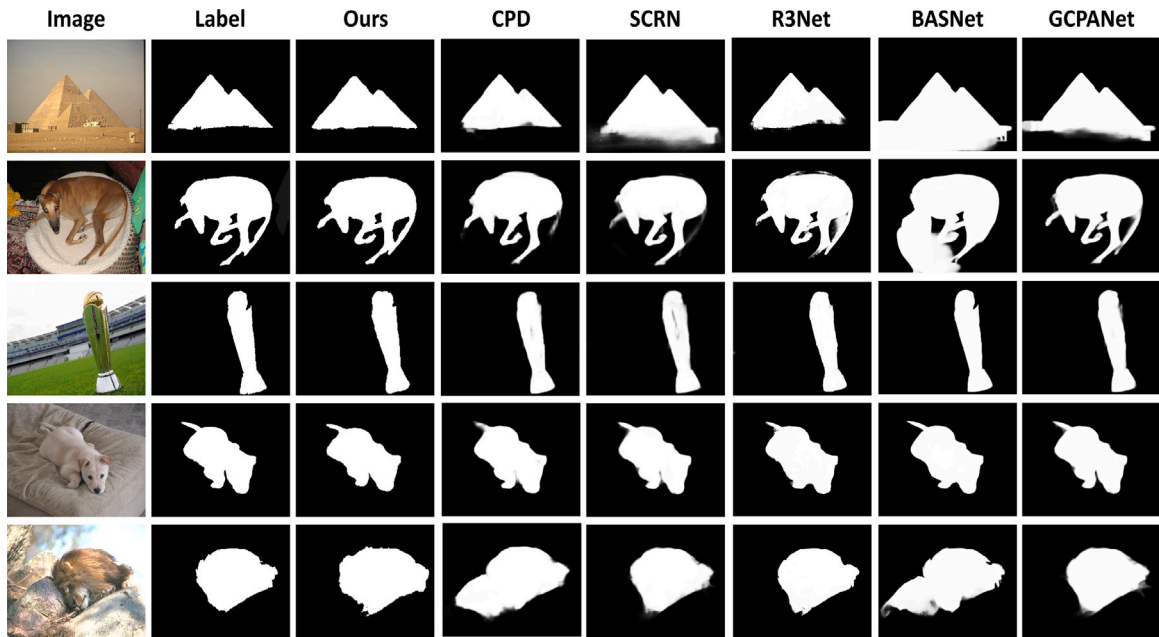


Fig. 8. Visual comparison of our framework predictions with the input images, GT, and the predictions of other SOTA networks CPD [75], SCRNet [76], R3Net [77], BASNet [31] and GCPANet [29].

Table 5

Comparative analysis with SOTA methods in terms of standard evaluation metrics of MAE and f_β measures.

Methods	ECSSD		HKU-IS		PASCAL-S		DUTS-TE	
	MAE	f_β	MAE	f_β	MAE	f_β	MAE	f_β
R3Net [77]	0.051	0.896	0.047	0.868	0.103	0.815	0.067	0.778
SCRNet [76]	0.037	0.927	0.034	0.907	0.066	0.863	0.040	0.856
GCPANet [29]	0.035	0.925	0.030	0.915	0.063	0.861	0.038	0.858
DFI [83]	0.035	0.929	0.031	0.911	0.066	0.867	0.039	0.857
GateNet [84]	0.040	0.925	0.033	0.909	0.071	0.858	0.040	0.855
AFNet [63]	0.042	0.915	0.036	0.899	0.071	0.849	0.046	0.834
RAS [23]	0.045	0.901	0.039	0.889	0.076	0.820	0.049	0.802
PoolNet [64]	0.039	0.923	0.032	0.908	0.076	0.856	0.040	0.851
CPD [75]	0.037	0.924	0.034	0.904	0.074	0.852	0.043	0.843
CAGNet [67]	0.037	0.931	0.030	0.919	0.069	0.864	0.040	0.858
PAGR [26]	0.061	0.902	0.048	0.888	0.095	0.832	0.055	0.822
BASNet [62]	0.037	0.927	0.032	0.911	0.079	0.842	0.048	0.842
ACFFNet [33]	0.032	0.935	0.026	0.923	0.059	0.867	0.033	0.874
MINet [82]	0.033	0.931	0.029	0.917	0.066	0.860	0.037	0.860
MFRNet (Ours)	0.030	0.936	0.023	0.929	0.058	0.878	0.031	0.875

CRedit authorship contribution statement

Habib Khan: Writing – original draft, Visualization, Methodology, Conceptualization. **Muhammad Talha Usman:** Writing – review & editing, Validation, Formal analysis. **Imad Rida:** Visualization, Investigation, Formal analysis. **JaKeoung Koo:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition.

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.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00240740) and also supported by the Gachon University research fund of 2023 (GCU-202400420001).

Data availability

<https://github.com/habib1402/MFRNet-SOD>.

References

- [1] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, *IEEE Trans. Pattern Anal. Mach. Intell.* 20 (11) (1998) 1254–1259.
- [2] X. Zhao, Y. Pang, L. Zhang, H. Lu, L. Zhang, Suppress and balance: A simple gated network for salient object detection, in: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, Springer, 2020, pp. 35–51.
- [3] B. Li, L. Tang, S. Kuang, M. Song, S. Ding, Toward stable co-saliency detection and object co-segmentation, *IEEE Trans. Image Process.* 31 (2022) 6532–6547.
- [4] A. Shokoufandeh, I. Marsic, S.J. Dickinson, View-based object recognition using saliency maps, *Image Vis. Comput.* 17 (5–6) (1999) 445–460.
- [5] H. Fang, S. Gupta, F. Iandola, R.K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J.C. Platt, et al., From captions to visual concepts and back, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1473–1482.
- [6] M.-M. Cheng, Q.-B. Hou, S.-H. Zhang, P.L. Rosin, Intelligent visual media processing: When graphics meets vision, *J. Comput. Sci. Tech.* 32 (2017) 110–121.

- [7] J.-Y. Zhu, J. Wu, Y. Xu, E. Chang, Z. Tu, Unsupervised object class discovery via saliency-guided multiple class learning, *IEEE Trans. Pattern Anal. Mach. Intell.* 37 (4) (2014) 862–875.
- [8] J. Liu, R. Dian, S. Li, H. Liu, SGFusion: A saliency guided deep-learning framework for pixel-level image fusion, *Inf. Fusion* 91 (2023) 205–214.
- [9] Q. Wang, S. Tang, D. Zhai, X. Hu, Saliency based object tracking in complex scenes, *Neurocomputing* 314 (2018) 132–142.
- [10] C. Sun, X. Wu, J. Sun, C. Sun, M. Xu, Q. Ge, Saliency-induced moving object detection for robust RGB-D vision navigation under complex dynamic environments, *IEEE Trans. Intell. Transp. Syst.* 24 (10) (2023) 10716–10734.
- [11] A. Saporta, X. Gui, A. Agrawal, A. Pareek, S.Q. Truong, C.D. Nguyen, V.-D. Ngo, J. Seekins, F.G. Blankenberg, A.Y. Ng, et al., Benchmarking saliency methods for chest X-ray interpretation, *Nat. Mach. Intell.* 4 (10) (2022) 867–878.
- [12] F. Lateef, M. Kas, Y. Ruichek, Saliency heat-map as visual attention for autonomous driving using generative adversarial network (GAN), *IEEE Trans. Intell. Transp. Syst.* 23 (6) (2021) 5360–5373.
- [13] R. Achanta, S. Hemami, F. Estrada, S. Susstrunk, Frequency-tuned salient region detection, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 1597–1604.
- [14] F. Perazzi, P. Krähenbühl, Y. Pritch, A. Hornung, Saliency filters: Contrast based filtering for salient region detection, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 733–740.
- [15] C. Yang, L. Zhang, H. Lu, X. Ruan, M.-H. Yang, Saliency detection via graph-based manifold ranking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3166–3173.
- [16] J. Zhang, K.A. Ehinger, J. Ding, J. Yang, A prior-based graph for salient object detection, in: 2014 IEEE International Conference on Image Processing, ICIP, IEEE, 2014, pp. 1175–1178.
- [17] M.-M. Cheng, N.J. Mitra, X. Huang, P.H. Torr, S.-M. Hu, Global contrast based salient region detection, *IEEE Trans. Pattern Anal. Mach. Intell.* 37 (3) (2014) 569–582.
- [18] N. Tong, H. Lu, X. Ruan, M.-H. Yang, Salient object detection via bootstrap learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1884–1892.
- [19] M. Munsif, S.U. Khan, N. Khan, A. Hussain, M.J. Kim, S.W. Baik, Contextual visual and motion salient fusion framework for action recognition in dark environments, *Knowl.-Based Syst.* (2024) 112480.
- [20] G. Lee, Y.-W. Tai, J. Kim, Deep saliency with encoded low level distance map and high level features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 660–668.
- [21] G. Li, Y. Yu, Visual saliency based on multiscale deep features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5455–5463.
- [22] T. Wang, A. Borji, L. Zhang, P. Zhang, H. Lu, A stagewise refinement model for detecting salient objects in images, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4019–4028.
- [23] S. Chen, X. Tan, B. Wang, X. Hu, Reverse attention for salient object detection, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 234–250.
- [24] T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, A. Borji, Detect globally, refine locally: A novel approach to saliency detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3127–3135.
- [25] P. Zhang, D. Wang, H. Lu, H. Wang, X. Ruan, Amulet: Aggregating multi-level convolutional features for salient object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 202–211.
- [26] X. Zhang, T. Wang, J. Qi, H. Lu, G. Wang, Progressive attention guided recurrent network for salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 714–722.
- [27] N. Liu, J. Han, M.-H. Yang, Picanet: Learning pixel-wise contextual attention for saliency detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3089–3098.
- [28] B. Wang, Q. Chen, M. Zhou, Z. Zhang, X. Jin, K. Gai, Progressive feature polishing network for salient object detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 12128–12135.
- [29] Z. Chen, Q. Xu, R. Cong, Q. Huang, Global context-aware progressive aggregation network for salient object detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 10599–10606.
- [30] T. Zhao, X. Wu, Pyramid feature attention network for saliency detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3085–3094.
- [31] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, M. Jagersand, Basnet: Boundary-aware salient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7479–7489.
- [32] M. Zhang, T. Liu, Y. Piao, S. Yao, H. Lu, Auto-msfnet: Search multi-scale fusion network for salient object detection, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 667–676.
- [33] J. Zhang, Y. Shi, Q. Zhang, L. Cui, Y. Chen, Y. Yi, Attention guided contextual feature fusion network for salient object detection, *Image Vis. Comput.* 117 (2022) 104337.
- [34] W. Zhu, S. Liang, Y. Wei, J. Sun, Saliency optimization from robust background detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2814–2821.
- [35] M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, N. Crook, Efficient salient region detection with soft image abstraction, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1529–1536.
- [36] R. Margolin, A. Tal, L. Zelnik-Manor, What makes a patch distinct? in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 1139–1146.
- [37] J. Shi, Q. Yan, L. Xu, J. Jia, Hierarchical image saliency detection on extended CSSD, *IEEE Trans. Pattern Anal. Mach. Intell.* 38 (4) (2015) 717–729.
- [38] Q. Xu, F. Wang, Y. Gong, Z. Wang, K. Zeng, Q. Li, X. Luo, A novel edge-oriented framework for saliency detection enhancement, *Image Vis. Comput.* 87 (2019) 1–12.
- [39] Y.-F. Ma, H.-J. Zhang, Contrast-based image attention analysis by using fuzzy growing, in: Proceedings of the Eleventh ACM International Conference on Multimedia, 2003, pp. 374–381.
- [40] M. Cerf, J. Harel, W. Einhäuser, C. Koch, Predicting human gaze using low-level saliency combined with face detection, *Adv. Neural Inf. Process. Syst.* 20 (2007).
- [41] R. Achanta, S. Hemami, F. Estrada, S. Susstrunk, Frequency-tuned salient region detection, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 1597–1604.
- [42] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, S. Li, Salient object detection: A discriminative regional feature integration approach, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2083–2090.
- [43] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, H.-Y. Shum, Learning to detect a salient object, *IEEE Trans. Pattern Anal. Mach. Intell.* 33 (2) (2010) 353–367.
- [44] J. Zhai, T. Litzhong, Y. Zhuang, Saliency detection based on boundary prior and adaptive region merging, *Comput. Eng. Appl.* 54 (6) (2018) 178–182.
- [45] H. Lu, X. Li, L. Zhang, X. Ruan, M.-H. Yang, Dense and sparse reconstruction error based saliency descriptor, *IEEE Trans. Image Process.* 25 (4) (2016) 1592–1603.
- [46] L. Zhang, J. Ai, B. Jiang, H. Lu, X. Li, Saliency detection via absorbing Markov chain with learnt transition probability, *IEEE Trans. Image Process.* 27 (2) (2017) 987–998.
- [47] L. Zhang, C. Yang, H. Lu, X. Ruan, M.-H. Yang, Ranking saliency, *IEEE Trans. Pattern Anal. Mach. Intell.* 39 (9) (2016) 1892–1904.
- [48] K.-Y. Chang, T.-L. Liu, H.-T. Chen, S.-H. Lai, Fusing generic objectness and visual saliency for salient object detection, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 914–921.
- [49] Y. Li, X. Hou, C. Koch, J.M. Rehg, A.L. Yuille, The secrets of salient object segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 280–287.
- [50] M. Munsif, N. Khan, A. Hussain, M.J. Kim, S.W. Baik, Darkness-adaptive action recognition: Leveraging efficient tubelet slow-fast network for industrial applications, *IEEE Trans. Ind. Inform.* (2024).
- [51] H. Yar, Z.A. Khan, I. Rida, W. Ullah, M.J. Kim, S.W. Baik, An efficient deep learning architecture for effective fire detection in smart surveillance, *Image Vis. Comput.* 145 (2024) 104989.
- [52] Z.A. Khan, T. Hussain, F.U.M. Ullah, S.K. Gupta, M.Y. Lee, S.W. Baik, Randomly initialized CNN with densely connected stacked autoencoder for efficient fire detection, *Eng. Appl. Artif. Intell.* 116 (2022) 105403.
- [53] A. Hussain, S.U. Khan, I. Rida, N. Khan, S.W. Baik, Human centric attention with deep multiscale feature fusion framework for activity recognition in internet of medical things, *Inf. Fusion* 106 (2024) 102211.
- [54] T. Khan, G. Choi, S. Lee, EFFNet-CA: an efficient driver distraction detection based on multiscale features extractions and channel attention mechanism, *Sensors* 23 (8) (2023) 3835.
- [55] M.T. Usman, H. Khan, S.K. Kumar, M.Y. Lee, J. Koo, Efficient deepfake detection via layer-frozen assisted dual attention network for consumer imaging devices, *IEEE Trans. Consum. Electron.* (2024).
- [56] W. Wang, J. Shen, X. Dong, A. Borji, Salient object detection driven by fixation prediction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1711–1720.
- [57] L. Wang, H. Lu, X. Ruan, M.-H. Yang, Deep networks for saliency detection via local estimation and global search, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3183–3192.
- [58] R. Zhao, W. Ouyang, H. Li, X. Wang, Saliency detection by multi-context deep learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1265–1274.
- [59] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, P.H. Torr, Deeply supervised salient object detection with short connections, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3203–3212.
- [60] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, M.-M. Cheng, EGNet: Edge guidance network for salient object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8779–8788.
- [61] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, European conference on computer vision, 2016.

- [62] J. Su, J. Li, Y. Zhang, C. Xia, Y. Tian, Selectivity or invariance: Boundary-aware salient object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 3799–3808.
- [63] M. Feng, H. Lu, E. Ding, Attentive feedback network for boundary-aware salient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1623–1632.
- [64] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, 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, 2019, pp. 3917–3926.
- [65] W. Wang, S. Zhao, J. Shen, S.C. Hoi, A. Borji, Salient object detection with pyramid attention and salient edges, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1448–1457.
- [66] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, P.-M. Jodoin, Non-local deep features for salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6609–6617.
- [67] S. Mohammadi, M. Noori, A. Bahri, S.G. Majelan, M. Havaei, CAGNet: Content-aware guidance for salient object detection, *Pattern Recognit.* 103 (2020) 107303.
- [68] M. Tan, Efficientnet: Rethinking model scaling for convolutional neural networks, 2019, arXiv preprint arXiv:1905.11946.
- [69] C. Peng, X. Zhang, G. Yu, G. Luo, J. Sun, Large kernel matters—improve semantic segmentation by global convolutional network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4353–4361.
- [70] M. Munsif, M. Sajjad, M. Ullah, A.N. Tarekegn, F.A. Cheikh, P. Tsakanikas, K. Muhammad, Optimized efficient attention-based network for facial expressions analysis in neurological health care, *Comput. Biol. Med.* 179 (2024) 108822.
- [71] H. Khan, I. Ullah, M. Shabaz, M.F. Omer, M.T. Usman, M.S. Guellil, J. Koo, Visionary vigilance: Optimized YOLOV8 for fallen person detection with large-scale benchmark dataset, *Image Vis. Comput.* 149 (2024) 105195.
- [72] Q. Yan, L. Xu, J. Shi, J. Jia, Hierarchical saliency detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 1155–1162.
- [73] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, X. Ruan, Learning to detect salient objects with image-level supervision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 136–145.
- [74] T. Khan, Z.A. Khan, C. Choi, Enhancing real-time fire detection: an effective multi-attention network and a fire benchmark, *Neural Comput. Appl.* (2023) 1–15.
- [75] Z. Wu, L. Su, Q. Huang, Cascaded partial decoder for fast and accurate salient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3907–3916.
- [76] Z. Wu, L. Su, Q. Huang, Stacked cross refinement network for edge-aware salient object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7264–7273.
- [77] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, P.-A. Heng, R3net: Recurrent residual refinement network for saliency detection, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence, Vol. 684690, AAAI Press Menlo Park, CA, USA, 2018.
- [78] L. Wang, L. Wang, H. Lu, P. Zhang, X. Ruan, Saliency detection with recurrent fully convolutional networks, in: *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, Springer, 2016, pp. 825–841.
- [79] G. Li, Y. Yu, Deep contrast learning for salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 478–487.
- [80] N. Liu, J. Han, Dhsnet: Deep hierarchical saliency network for salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 678–686.
- [81] P. Zhang, D. Wang, H. Lu, H. Wang, B. Yin, Learning uncertain convolutional features for accurate saliency detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 212–221.
- [82] Y. Pang, X. Zhao, L. Zhang, H. Lu, Multi-scale interactive network for salient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9413–9422.
- [83] J.-J. Liu, Q. Hou, M.-M. Cheng, Dynamic feature integration for simultaneous detection of salient object, edge, and skeleton, *IEEE Trans. Image Process.* 29 (2020) 8652–8667.
- [84] X. Zhao, Y. Pang, L. Zhang, H. Lu, L. Zhang, Suppress and balance: A simple gated network for salient object detection, in: *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, Springer, 2020, pp. 35–51.
- [85] Y.-H. Wu, Y. Liu, L. Zhang, M.-M. Cheng, B. Ren, EDN: Salient object detection via extremely-downsampled network, *IEEE Trans. Image Process.* 31 (2022) 3125–3136.