TYPE Original Research
PUBLISHED 14 August 2024
DOI 10.3389/fcomp.2024.1420965

EDITED BY
Jing Yuan,
Zhejiang Normal University, China

REVIEWED BY
Jinhua Xu,
Zhejiang Normal University, China
Muhammad Aamir,
Huanggang Normal University, China

*CORRESPONDENCE
Ravi Kant Kumar
✉ ravikant.k@srmap.edu.in

RECEIVED 21 April 2024
ACCEPTED 15 July 2024
PUBLISHED 14 August 2024

CITATION
Dhara G and Kumar RK (2024) Spatial attention guided cGAN for improved salient object detection. *Front. Comput. Sci.* 6:1420965.
doi: 10.3389/fcomp.2024.1420965

# Spatial attention guided cGAN for improved salient object detection

Gayathri Dhara and Ravi Kant Kumar*

Department of Computer Science and Engineering (CSE), SRM University, Amaravathi, Andhra Pradesh, India

Recent research shows that Conditional Generative Adversarial Networks (cGANs) are effective for Salient Object Detection (SOD), a challenging computer vision task that mimics the way human vision focuses on important parts of an image. However, implementing cGANs for this task has presented several complexities, including instability during training with skip connections, weak generators, and difficulty in capturing context information for challenging images. These challenges are particularly evident when dealing with input images containing small salient objects against complex backgrounds, underscoring the need for careful design and tuning of cGANs to ensure accurate segmentation and detection of salient objects. To address these issues, we propose an innovative method for SOD using a cGAN framework. Our method utilizes encoder-decoder framework as the generator component for cGAN, enhancing the feature extraction process and facilitating accurate segmentation of the salient objects. We incorporate Wasserstein-1 distance within the cGAN training process to improve the accuracy of finding the salient objects and stabilize the training process. Additionally, our enhanced model efficiently captures intricate saliency cues by leveraging the spatial attention gate with global average pooling and regularization. The introduction of global average pooling layers in the encoder and decoder paths enhances the network's global perception and fine-grained detail capture, while the channel attention mechanism, facilitated by dense layers, dynamically modulates feature maps to amplify saliency cues. The generated saliency maps are evaluated by the discriminator for authenticity and gives feedback to enhance the generator's ability to generate high-resolution saliency maps. By iteratively training the discriminator and generator networks, the model achieves improved results in finding the salient object. We trained and validated our model using large-scale benchmark datasets commonly used for salient object detection, namely DUTS, ECSSD, and DUT-OMRON. Our approach was evaluated using standard performance metrics on these datasets. Precision, recall, MAE and $F\beta$ score metrics are used to evaluate performance. Our method achieved the lowest MAE values: 0.0292 on the ECSSD dataset, 0.033 on the DUTS-TE dataset, and 0.0439 on the challenging and complex DUT-OMRON dataset, compared to other state-of-the-art methods. Our proposed method demonstrates significant improvements in salient object detection, highlighting its potential benefits for real-life applications.

KEYWORDS

**computer vision, Conditional Generative Adversarial Networks (cGANs), context information, encoder-decoder framework, image segmentation, salient object detection, spatial attention**

## 1 Introduction

The concept of visual attention has motivated numerous researchers to replicate the human visual system's capabilities in computer vision. SOD is a technique that seeks to replicate the attention mechanisms observed in human vision. It enables the identification of prominent areas in an image that capture human gaze. Within the realm of computer vision, saliency detection serves as a cornerstone for many of the applications centered around image comprehension. These applications encompass diverse areas, including image compression (Zünd et al., 2013), scene classification (Qi and Wang, 2016), object localization (Aamir et al., 2023), object tracking (Borji and Itti, 2012), and various multimedia applications (Singh, 2020; Sun et al., 2022). The U-Net (Ronneberger et al., 2015), a representative example of a fully convolutional neural network, has achieved remarkable success in medical image segmentation. Its effectiveness in suppressing background noise is achieved through a two-step process: jointly employing an encoder and decoder to process image data, and then integrating the information using skip connections. When utilized for the purpose of saliency detection in natural images, the challenges become more intricate. This complexity arises from the presence of objects that are challenging to differentiate from complex backgrounds. These backgrounds include elements like pixel blocks with substantial contrast variations and mirrored reflections of salient objects (Zhang et al., 2019). Therefore, an enhanced methodology is imperative to effectively address the difficulties in distinguishing salient objects from intricate backgrounds.

This paper proposes a novel encoder-decoder network with attention mechanisms to address the limitations of existing methods in SOD. SOD aims to identify and segment the most visually important objects in an image. We formulate this task as a binary segmentation problem, where the goal is to distinguish the salient foreground objects from the non-salient background.The resulting segmentation map represents a binary mask, where the foreground regions are typically assigned as white or 1, while the background regions are marked as black or 0. Binary segmentation has numerous applications in vision computing, which include object detection, image editing, medical image analysis, and scene understanding. It provides a foundation for more complex segmentation tasks and enables subsequent analysis and processing of the segmented regions. In the past few years, cGANs have evolved as a robust framework for various tasks image generation, style transfer, and semantic segmentation. Numerous studies have been dedicated to enhancing the theoretical foundations of GANs (Goodfellow et al., 2014; Gulrajani et al., 2017; Isola et al., 2017), Goodfellow et al. (2014) introduced the original GAN model, which aimed to produce synthetic samples that are indistinguishable from real data by training a generative model in an adversarial fashion. In this model, the discriminator is tasked with discerning between fake and real samples, while the generator is trained to deceive the discriminator. The conditional nature of cGANs allows for the generation of outputs conditioned on input data, making them specifically well-suited for applications where the output is dependent on specific input information. The architectural difference between GAN and cGAN is as shown in Figure 1.

The motivation to adopt our proposed approach resides in its skill to manage the limitations of existing U-Net-based methods in salient object detection. By incorporating attention mechanisms, we aim to enhance model's capability to emphasize relevant spatial regions while selectively suppressing irrelevant or distracting elements. This attention-guided mechanism is expected to significantly improve the accuracy and robustness of salient object detection, leading to more precise and reliable segmentation results.

The main contributions of this research are:

- To investigate the application of cGAN with an enhanced encoder-decoder network architecture in the context of SOD. We put forward an enhanced encoder-decoder framework that strategically addresses the differences between abstract and detailed features while incorporating the unique characteristics of the contracting path and expanding path.
- Attention mechanisms are incorporated to address the difficulty of SOD by selectively amplifying the significance of prominent spatial regions and channels within the input data. This focused approach allows the network to extract more meaningful features from these regions, leading to improved accuracy and precision in tasks like segmentation and detection.
- To comprehensively evaluate our method's performance, we conducted experimental analysis on three challenging datasets for SOD. The results convincingly demonstrate the performance of our proposed network in this task.

The related work is given as part of Section 2, and the proposed method for SOD is introduced and elaborated under Section 3. Details of metrics used for evaluation are part of the Section 4.2. The results and discussion are detailed in Section 4. The quintessence of the proposed work is summarized in Section 6.

## 2 Related works

Drawing inspiration from Feature Integration Theory (FIT) (Treisman and Gelade, 1980), Koch and Ullman (1987) developed a computational architecture grounded in biological plausibility to model human selective attention mechanisms by replicating the early feature representations described in FIT.

In practice, the development of bottom-up methods significantly depends on various saliency priors such as center-surround prior, foreground prior, boundary connectivity prior, local and global contrast prior, focusness prior, and geodesic prior (Perazzi et al., 2012; Wei et al., 2012; Jiang et al., 2013a; Yang et al., 2013; Cheng et al., 2014; Li et al., 2014). These priors function as semi-supervised guides, influencing the creation of heuristic bottom-up techniques constrained by these factors. For example, Zhu et al. (2014) presented a saliency optimization approach that calculates the background probability of superpixels using geodesic saliency. Likewise, Yang et al. (2013) proposed a graph-based model with manifold ranking, where boundary nodes are generally treated as background or non-salient. However, bottom-up methods are fundamentally heuristic and dependent on specific saliency priors, which can limit their effectiveness when images do not conform

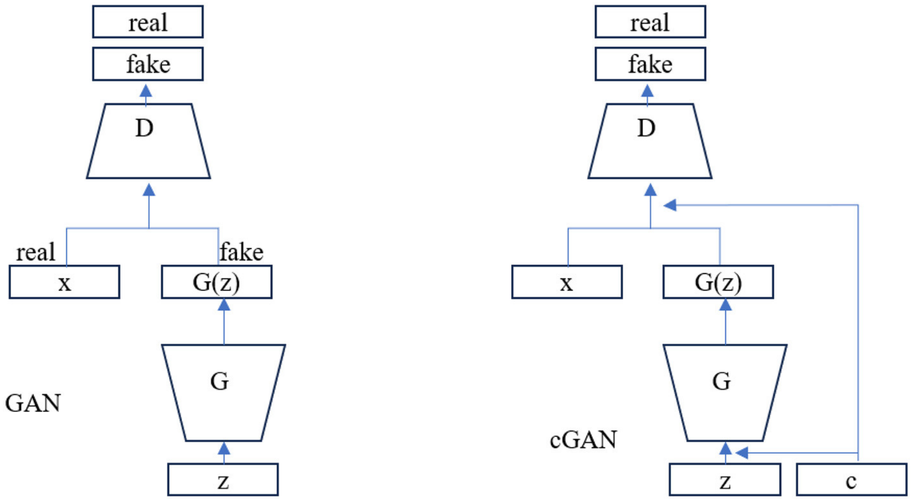

FIGURE 1
Architectural difference between GAN and cGAN Mino and Spanakis (2018).

well to these priors. This dependence on unsupervised techniques based on preset constraints creates a performance bottleneck. In contrast, top-down approaches focus more on feature extraction and classifier design, especially before the deep learning era. Jiang et al. (2013b) proposed a discriminative regional feature integration approach that uses a random forest regressor to map regional feature vectors to saliency scores. A method for detecting salient objects by modeling saliency features through conditional random field (CRF) learning in (Liu et al., 2010).

**Region proposal algorithms** lay the groundwork for faster Recurrent Neural Network (R-CNN) models in object detection. These algorithms slide a compact network over a feature map, classifying objects within specific regions. Girshick et al. (2014) suggested the region-based convolution neural network (R-CNN) to address the challenge of region generation, leveraging the selective search algorithm. Introducing Faster R-CNN, Ren et al. (2015) revolutionized neural networks by eliminating the selective search algorithm for region proposal extraction. Instead, they harnessed a separate network, the Region Proposal Network (RPN), to predict region proposals, resulting in a substantial performance boost. Using RPN for region generation enables Faster R-CNN to outpace the execution speed of R-CNN and fast R-CNN. For precise pixel-level image segmentation, Buric et al. (2018) put forth Mask R-CNN, a network incorporating aligned Region of Interest (ROI) Pooling, yielding enhanced accuracy in pixel-wise image segmentation. Nonetheless, this network does suffer from slower detection times. Overall, the evolution from R-CNN to Fast R-CNN and finally to Faster R-CNN has focused on overcoming the limitations of the selective search algorithm while improving the efficiency and precision of object detection. (Guan et al., 2021) introduced a method for image object detection and classification utilizing a deep neural network (DNN) that relies on precise object localization.

**Deep convolutional neural networks (CNNs)** have brought about remarkable enhancements in image classification and object detection when contrasted with manually crafted features. These CNN-based approaches adeptly extract semantic information or features at multiple levels and integrate them with high efficiency. Traditional methods for salient object detection primarily depend on intrinsic cues to generate saliency maps. However, these methods were constrained in their capacity to extract high-level semantic features. The work of Chen et al. (2014) gave rise to DeepLab, an architecture that integrates deep CNNs, substituting conventional downsampling with atrous convolutions within convolutional layers. Expanding upon the foundations of DeepLab, Chen et al. (2017b) introduced DeepLabV3, surpassing its predecessors, DeepLabV1 and DeepLabV2, through heightened segmentation task efficiency. This network was meticulously designed to yield feature maps that adeptly encapsulate multiscale content. In recent developments, Chen et al. (2018) introduced DeepLabV3+, an extension of DeepLabV3. This model integrates the Xception Model and applies depth-wise separable convolutions to both the atrous spatial pyramid pooling (ASPP) and decoder modules, yielding a discernible advancement in performance. The introduction of a multi-level feature aggregation network, AmuletNet, which utilizes convolutional features from multiple levels as saliency cues for salient object detection is proposed by Zhang et al. (2017). The design of a network module that can capture and aggregate global context information from different scales and levels, which progressively refine the saliency prediction, is proposed in Chen et al. (2020).

**Methods based on upsampling/deconvolution** have also been suggested for image segmentation. In this context, Jégou et al. (2017) introduced FC-DenseNet, a model built upon the DenseNet architecture. FC-DenseNet harnesses pre-trained parameters and employs post-processing techniques to enhance scene comprehension. By incorporating both upsampling and downsampling pathways alongside skip connections, FC-DenseNet effectively retains and utilizes spatial information.

Badrinarayanan et al. (2017) suggested SegNet for precise pixel-wise classification and boundary localization. This approach employs an encoder-decoder network structure, with each encoder layer correspondingly linked to its decoder counterpart. However, the computational intricacy associated with pixel-wise labeling presents challenges for real-time scene segmentation. Additionally, SegNet's prolonged inference time makes it less suitable for applications requiring rapid responses. In response to these limitations, Zhao et al. (2018) presented ICNet, a network designed to attain swift and accurate segmentation. These methods based on upsampling/deconvolution, encompassing FC-DenseNet, SegNet, and ICNet, have notably advanced semantic segmentation by enabling meticulous pixel-wise classification and precise boundary delineation. While each approach boasts distinct merits and drawbacks, they collectively offer valuable insights for addressing challenges in scene comprehension and real-time segmentation.

**Feature Encoder-Based Methods:** These methods rely on converting categorical variables into continuous ones for effective integration into the model. For instance, VGG Net was introduced by Simonyan and Zisserman (2015), which incorporated a series of consecutive 3×3 convolutions inspired by the architecture of AlexNet (Krizhevsky et al., 2012). However, as the network's depth increased, performance saturation and degradation challenges emerged. To tackle these issues, He et al. (2016) introduced ResNet (residual network), which addressed the problem of vanishing gradients by incorporating residual blocks in a pre-activation configuration. While effective, ResNet's dense residual layers led to heightened computational complexity, especially when dealing with datasets of limited size. In recent years, many deep CNN-based models have adopted an encoder-decoder framework based on FCN and can be trained end-to-end using pixel-wise annotated saliency maps. Specifically, these models typically operate within an FCN-like structure, originally designed for other image-to-image learning tasks such as semantic segmentation (Chen et al., 2017a, 2018) and edge detection (Xie and Tu, 2015). Techniques like skip connections (Ronneberger et al., 2015), atrous convolution (Chen et al., 2017a), and pyramid pooling modules (Zhao et al., 2017) can be implicitly or explicitly incorporated to develop new deep CNN models for SOD. Recently, several multi-scale techniques have been developed to enhance the ability to learn semantic information. For instance, MINet (Pang et al., 2020) is a network that merges features from adjacent layers and detects salient objects by minimizing noise from resolution differences in feature maps using small up-/down-sampling rates. Chen et al. (2019) proposed an innovative method featuring a parallel multi-scale structure to integrate salient features at various levels. Ji et al. (2018b) introduced a technique for learning context between feature information of different scales by applying spatial and channel attention modules to multi-scale encoder-decoder networks. To tackle the challenge of scale variations in objects, most SOD models focus on enhancing multi-scale feature fusion by incorporating advanced feature fusion modules within their networks. ICON (Zhuge et al., 2022) incorporated convolutional kernels of various shapes to boost feature diversity through multi-level feature fusion. MENet (Wang et al., 2023) utilized atrous spatial pyramid pooling (ASPP) (Lian et al., 2021) to combine multi-scale features, thereby improving feature representation. Additionally, several techniques have been developed to enhance segmentation performance in salient object detection (SOD) by incorporating contour or edge information from input images and prediction maps. For instance, the authors of (Zhao et al., 2019) introduced EGNet, which learns edge information via supervised learning using multiple branches, integrating local edge and global location information. Han et al. (2019) enhanced U-Net by adding an edge convolution constraint to achieve more accurate saliency maps. Within the current body of literature, the significance of computational time in deep learning-based segmentation techniques has garnered recognition. Nonetheless, it has been noted that enhancements in a model's performance are not necessarily directly proportional to the computational time invested. Moreover, diverse models face specific challenges and constraints. For instance, feature-encoder-based models often grapple with memory limitations, particularly when applied to expansive datasets (Simonyan and Zisserman, 2015). Existing deep learning methods for salient object detection have made significant advancements, but they still suffer from several drawbacks, such as overfitting, sensitivity to hyperparameters, real-time performance issues, and model complexity and size. Our proposed architecture effectively addresses several issues by employing a conditional Generative Adversarial Network (cGAN) combined with an enhanced UNET architecture. The cGAN framework, through its adversarial training process, improves the model's ability to generalize across diverse datasets and conditions, enhancing robustness to noise and distortions. The enhanced UNET architecture, known for its efficient use of convolutional and deconvolutional layers, reduces computational complexity and achieves finer granularity in segmentation tasks. This combination not only improves real-time performance but also ensures precise boundary detection and better interpretability of results. Since the generator is trained using the pre-trained weights of an enhanced UNET, the convergence rate is significantly improved.

# 3 Proposed method

## 3.1 Method overview

Our proposed model, depicted in Figure 2, is designed as a combination of two networks: a generative network and a discriminator network. The generative network utilizes an encoder-decoder mechanism, which learns the structural information of salient objects through adversarial learning. It is trained in an end-to-end fashion with a discriminator network to distinguish real salient maps from the fake ones produced by the generator. In this network design, we treated saliency detection as a semantic segmentation problem, as both tasks involve understanding and identifying essential regions in an image. SOD focuses on finding the most visually salient objects in an image, akin to the objective of object localization in semantic segmentation. Furthermore, both saliency detection and semantic segmentation take into account the relevant details of an image. In saliency detection, context is crucial for determining the relative importance of regions within an image. Similarly, semantic segmentation relies on contextual information to accurately assign labels to pixels based on their surrounding regions. The U-Net framework has demonstrated remarkable potential in various image analysis tasks, including semantic segmentation. However, its direct application to visual saliency detection may only partially leverage the unique characteristics of salient regions.

Furthermore, the original U-Net framework suffers from the following drawbacks: Firstly, while skip connections aid in the efficient transmission of low-resolution information in feature maps, they often lead to a blurring effect on the obtained image features. Additionally, the complex features that the network extracts often miss out on high-resolution edge details from the input. This lack of detail can cause uncertainty, especially in situations where high-resolution edges play a crucial role in the network's decisions, such as when detecting small or intricate salient objects like butterflies or flowers. In these cases, high-resolution edge information is vital for accurate segmentation. Lastly, finding the optimal frequency of pooling operations for extracting high-level global features presents a significant challenge, as the utilization of pooling operations is subject to change based on the size of the object. To address these issues, we established an encoder-decoder structure, as illustrated in Figure 3. Our approach integrates attention mechanisms into the encoder and decoder design to improve feature extraction and highlight salient regions. The encoder portion obtains multiscale features from the input image, while the decoder part recreates the saliency map by combining the encoded features with the attention-guided information. Our method utilizes spatial attention gate, an approach employed in CNNs to enhance their performance in segmenting an image and detection of object. This method selectively highlights important areas within feature maps, directing the network to concentrate on the most relevant sections of the input image. During the processing of the input feature map, a 1×1 convolution is utilized, which functions as a filter to assess the value of each pixel across all channels. The resulting output from this operation consists of a single channel that assigns an "attention score" to each individual pixel. These scores undergo a sigmoid activation function, transforming them into values ranging between 0 and 1. The attention scores essentially serve as a filter implemented individually to each element of the original input feature map. Pixels receiving high attention scores, nearing 1, undergo a boost, heightening their impact on the final result. Conversely, pixels with low attention scores, nearing 0, are dampened, decreasing their significance. In our proposed approach, we integrate the encoder-decoder architecture as the generator component of cGAN for generating salient maps. The benefits of attention mechanism followed are :

- **Improved performance:** by concentrating on significant areas, the network has the capability to extract more meaningful characteristics, resulting in improved accuracy and precision for tasks like segmentation and detection.
- Enhanced interpretability: the attention maps created by the gate can be displayed visually, offering insights into the image regions that the network deems most crucial for its decision-making process.
- Reduced Sensitivity to Noise: By suppressing irrelevant regions in the feature maps, the network becomes less vulnerable to noise and imperfections in the input image.

The model extracts features at various scales using convolutional layers, downsampling through pooling, and encoding spatial information. At a bottleneck point in pool4, the attention gate refines feature maps. In our enhanced U-Net framework, the gate is strategically positioned before the expansive path, ensuring attention-guided feature recovery.The upsampling process incorporates refined features, enabling attention-informed reconstruction. The gate produces spatially weighted features, guiding subsequent layers.

The architecture, trained using the DUTS-TR dataset, generates segmentation masks that accurately identify and emphasize the salient regions within the input images. On the otherhand, the discriminator component of the cGAN also receives the original input images along with their segmentation masks as part of the conditioning. cGANs consist of a generator and a discriminator, engaged in a competitive game aiming to enhance the quality of generated samples through a min-max framework. By incorporating conditional information, cGANs can generate images conditioned on specific attributes or input data. They learn to differentiate between actual saliency maps (generated from the segmentation masks and input images) and fake saliency maps (generated by the generator component). The discriminator uses the conditioning information to guide its discrimination process and provide feedback to the generator to enhance the saliency map generation. Specifically, within the cGAN framework, we utilize the Wasserstein distance to maintain the process of training. Additionally, an L2 norm loss is applied between the produced saliency mask and the ground-truth map. Furthermore, recent studies have shown that DCNN models (Zheng et al., 2015; Chen et al., 2017a), when combined with fully-connected conditional random fields (CRFs), can significantly improve semantic segmentation accuracy. This research encourages us to enhance the generated saliency mask by applying dense CRF inference as a post-processing step. The research works of Arjovsky et al. (2017); Ji et al. (2018a) incorporated the Earth-Mover (EM) distance into the traditional Generative Adversarial Network (GAN) model, addressing a fundamental challenge in the objective function of conventional GANs. The EM distance demonstrates robust performance, particularly in scenarios where two distributions do not overlap, distinguishing itself from other probability distances and divergences. Previous research has shown that this approach significantly enhances the stability of the training process and helps mitigate issues such as mode collapse. The initial training phase of the cGAN model is crucial for achieving a good detection model. The key to deceiving the discriminator during training is the adaptive feedback mechanism. When salient and non-salient samples can be easily distinguished, the discriminator becomes increasingly stronger throughout the training process. However, this could cause the generator into a significant mode collapse issue by producing negative samples with a limited pattern. To overcome this challenge, we introduce a technique that penalizes the generation of unrealistic images during training. This method compares the generated image to a real image from the target domain and focuses on minimizing the difference between them. Importantly, since the source image stays the same throughout the process, we only compare the generated image to the target image's true counterpart.

To train our cGAN model effectively, we define an objective function that specifies as

$$G^{*} = \arg\min_{G}\max_{D} L_{cGANw}(G, D) + \lambda L_{L1}(G) \tag{1}$$

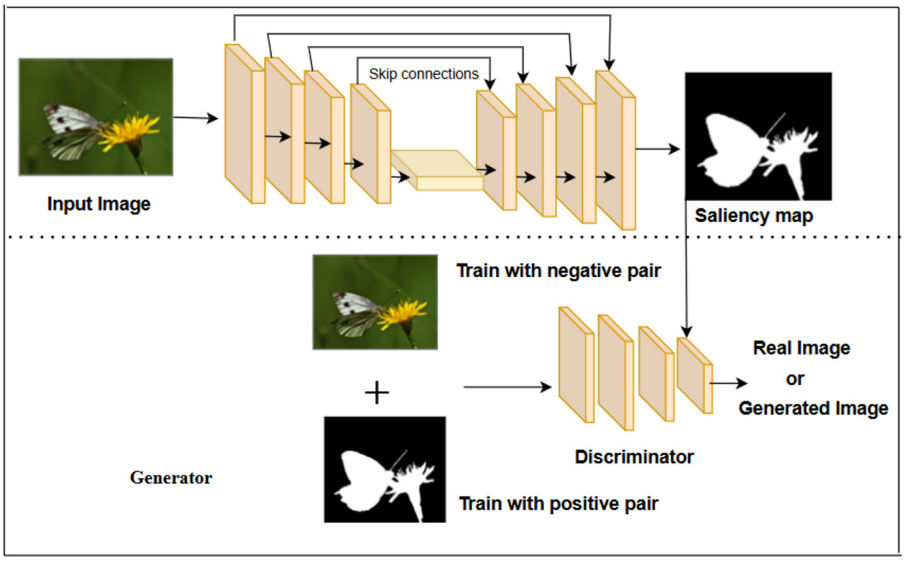

FIGURE 2
Proposed architecture. Input image is taken from the publicly available ECSSD dataset (image no: 0213).

In this equation, λ serves as a penalty weight for the loss between the ground-truth saliency and the generated saliency map. We assign a substantial weight to the L2-norm loss function to encourage the production of more challenging negative samples. Within this cGAN framework, the training process is guided by images and their corresponding ground-truth saliency maps.

Recent research (Chen et al., 2017a; Ji et al., 2018a) has shown significant improvements in semantic segmentation accuracy by augmenting CNNs with fully connected CRFs. Inspired by the success of densely connected Conditional Random Fields (CRFs) in improving pixel-level annotation tasks, we integrated this approach into our saliency mask generation process. This incorporation aims to further refine the generated saliency masks. Traditional CRFs employ sparse connections between nodes, hindering their ability to capture long-range dependencies within the image. To overcome this limitation, researchers have proposed fully connected CRF models, which establish connections between all node pairs. This model aims to overcome the sparse connectivity issue by introducing a comprehensive network of connections, enabling effective modeling of long-range dependencies. The energy function for the densely connected Conditional Random Field (CRF) model is expressed as:

$$E(y) = \sum_{i} \psi_u(y_i) + \sum_{i<j} \psi_p(y_i, y_j) \tag{2}$$

where y represents the variable, and the potential functions $\psi_u$ and $\psi_p$ operate on individual variables and pairs of variables, respectively. In Krähenbühl and Koltun (2011), a mean-field approximation using weighted Gaussians to model pairwise potential was introduced for efficient inference. The pairwise potential is specified as:

$$\psi_p(u_i, u_j) = \mu(u_i, u_j)\left[ w_1 \exp\left( -\frac{\|\mathbf{pos}_i - \mathbf{pos}_j\|^2}{2\sigma_\alpha^2} \right) - \frac{\|\mathbf{color}_i - \mathbf{color}_j\|^2}{2\sigma_\beta^2} + w_2 \exp\left( -\frac{\|\mathbf{pos}_i - \mathbf{pos}_j\|^2}{2\sigma_\gamma^2} \right) \right] \tag{3}$$

The unique modifications and improvements that make it exceptionally suitable for detecting prominent objects set the proposed method apart from the conventional U-Net design are:

The proposed architecture, composed of an encoding path and a decoding path, is adept at detecting prominent objects. The encoder, the initial segment of the network, captures, and abstracts meaningful features from the input data through several convolutional layers, each followed by batch normalization and activation functions. These layers operate on the input image, progressively extracting hierarchical features. The spatial dimensions of the feature maps are gradually reduced using max-pooling layers. A critical component, the bottleneck, compresses the learned features into a compact representation through global average pooling and dense (fully connected) layers. The decoder reconstructs the original spatial dimensions from the compressed representation and generates the final output. It consists of transposed convolutional layers and skip connections from the encoder. The final layers of the decoder involve additional convolutional operations, batch normalization, activation functions, and a convolutional layer with a single channel to generate the reconstructed output, matching the original input size of 256×256 pixels with a single channel. This method introduces several enhancements to the conventional U-Net design, making it particularly adept at detecting prominent objects. It incorporates dense layers and multiplication operations, which capture complex relationships within the data, thereby improving segmentation

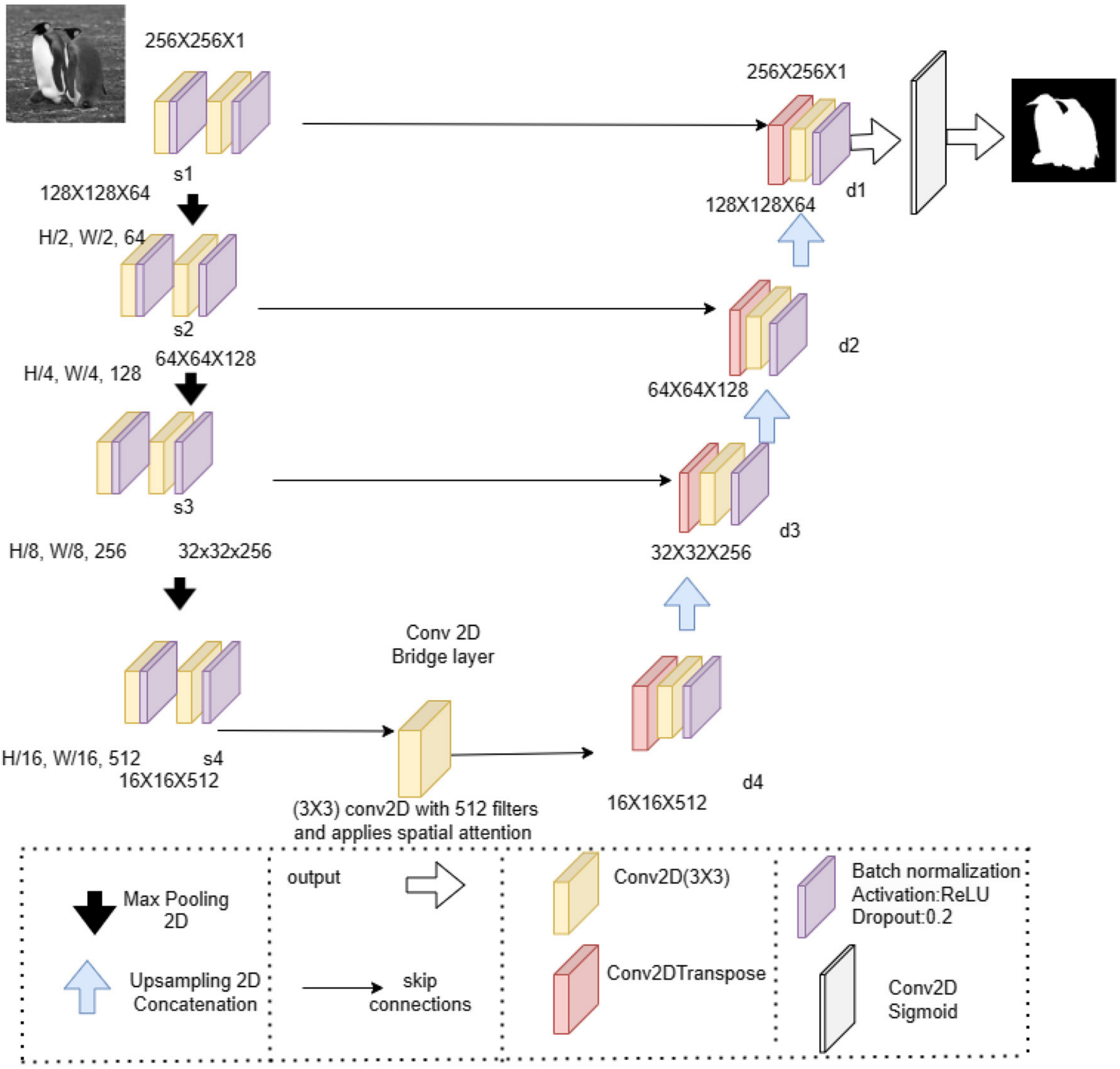

FIGURE 3
Detailed view of generator. Input image is taken from the DUTS-TR dataset (image no: ILSVRC2012_test_00000181).

accuracy. The model also extracts adaptive contextual information, which aids in segmenting intricate structures. A unique feature of this method is the hierarchical feature fusion mechanism, which emphasizes features based on their importance. The architecture is tailored to address specific segmentation challenges, demonstrating its adaptability. Integrating channel and spatial attention mechanisms enhances feature relevance and contextual awareness. These mechanisms, along with multi-layer feature fusion and adaptive contextual information integration, contribute to improved segmentation accuracy. This method marks a notable advancement in the realm of object detection and segmentation.

This enhanced U-Net variant showcases its distinctive strengths in scenarios where fine-grained segmentation, detailed feature extraction, and intricate boundary detection are paramount. Its hybrid approach, combining U-Net's foundational architecture with supplementary components, positions it as a versatile tool for applications ranging from medical image analysis to remote sensing. The proposed architecture is illustrated in Figure 2, and the detailed generator module is provided under Figure 3. The discriminator receives the actual input image, its corresponding mask, and the saliency map generated by the generator. These inputs are fed into the discriminator network for analysis and classification. The primary function of the discriminator is to assess the validity of the generated saliency map, learning to differentiate between genuine saliency maps and those produced by the generator. The discriminator architecture as shown in Figure 4 consists of four convolutional layers, each with a 3×3 kernel size, followed by a Leaky-ReLU activation layer and a max-pooling layer. These additional layers help extract relevant features and downsample the data, enabling the discriminator to learn hierarchical representations of the input data. The convolutional layers capture discriminative features at multiple scales, while the Leaky-ReLU activation function aids gradient flow and prevents neuron saturation, ensuring effective information propagation within the network. The max-pooling layers reduce the spatial dimensions, providing a compressed representation of the features obtained by the convolutional layers. In the final step of the model, a sigmoid activation function is applied to the last layer. This function produces a score that indicates whether the saliency map originates from the generator or the ground truth. The sigmoid function outputs values between 0 and 1, allowing for a probabilistic interpretation of the score. A score close to 1 suggests a high probability that the saliency map corresponds to the ground truth, while values closer to 0 indicate that the generator produced the saliency map. During the training process, the discriminator aims to accurately classify input saliency maps as real or generated, minimizing classification error and improving its ability to distinguish between the two types of maps. As training progresses, the discriminator becomes more skilled at identifying subtle differences between real and generated saliency maps. This feedback is crucial for the generator's learning process, enabling it to refine its saliency map generation by producing maps that are harder for the discriminator to classify as generated. The adversarial interplay between the generator and discriminator networks helps the generator produce saliency maps that are both visually accurate and realistic. The discriminator uses binary cross-entropy loss to classify the real and fake pairs.

In summary, the encoder, bottleneck, and decoder work together to transform the input image into a compressed representation, process it to learn relevant features, and finally decode the representation to generate the final output. The architecture's sizes and parameters, such as filter sizes, channel counts, and spatial dimensions, play a crucial role in its ability to extract and manipulate intricate image features. To enhance the performance of cGAN, several strategies were employed in this study, as outlined in Mino and Spanakis (2018). These included incorporating batch normalization, using leakyReLU activation, initializing weights randomly from a Gaussian distribution with a standard deviation of 0.002 for both the discriminator and generator. Additionally, L2 regularization was applied to all layers of the discriminator. These changes significantly improved the convergence speed of the model. The sizes of all feature vectors in the input and output layers of the generator are illustrated in Figure 3.

## 3.2 Objective function

The objective function of conditional GAN can be expressed as:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x), y\sim p_{\text{data}}(y)}[\log D(x,y)] + \mathbb{E}_{z\sim p_z(z), y\sim p_{\text{data}}(y)}[\log(1 - D(G_{\text{U-Net}}(z,y),y))] \tag{4}$$

The training process involves two main goals, represented by the loss function. The first part focuses on the discriminator, aiming to improve its ability to distinguish real data (like real photos) from generated data (like artificial images). The second part helps the generator create more realistic data by penalizing it when the discriminator correctly identifies the produced data as fake.

## 3.3 Loss functions

**BCE Loss**: We have used the BCE loss function in training our enhanced U-Net model to classify the foreground and background.

$$L_{BCE} = -\sum_{i=1}^{H}\sum_{j=1}^{w}\left[G_{ij} log P_{ij} + \left(1 - G_{ij}\right) log\left(1 - P_{ij}\right)\right] \tag{5}$$

**Generator Loss** The generator's loss function plays an important role and consists of two main components:

- **Adversarial Loss** :this component aims to fool the discriminator by minimizing the difference between the discriminator's predictions on the generated images and the true labels.
- **MSE loss**: this standard loss function measures the average squared differences between the generated output and the target output. Also known as L2 loss.

The total generator loss is given by:

$$L_{\text{total}} = L_{\text{adv}} + \lambda_{\text{MSE}} \times \text{MSE Loss} \tag{6}$$

Where: $L_{\text{total}}$ is the total generator loss. $L_{\text{adv}}$ is the adversarial loss. $\lambda_{\text{MSE}}$ is a hyperparameter controlling the relative importance of the MSE Loss.

The adversarial loss is computed by evaluating the probabilities assigned by the discriminator $D$ to the reconstructed images $G(I)$ across all training samples:

$$L_{\text{adv}} = -\sum_{n=1}^{N} \log D(G(I_n)) \tag{7}$$

The MSE Loss measures the pixel-wise difference between the generated images and the ground truth images:

$$\text{L2 Loss} = \frac{1}{N}\sum_{i=1}^{N}(I_{\text{GT}_i} - G(I_i))^2 \tag{8}$$

**Discriminator Loss**: The overall discriminator loss comprises the real data loss and the fake data loss. We use the BCE Loss (binary cross-entropy loss) function to calculate it:

$$\mathcal{L}_{\text{D}} = -\left(\frac{1}{N}\sum_{i=1}^{N}\log\left(D(x_i)\right) + \frac{1}{N}\sum_{i=1}^{N}\log\left(1 - D(G(z_i))\right)\right) \tag{9}$$

Where: - $D(x_i)$ represents the discriminator's output for a real data sample $x_i$. - $D(G(z_i))$ represents the generator's output for a noise sample $z_i$. - $N$ is the batch size. - log denotes the natural logarithm.

The first term computes the average binary cross-entropy loss for real data samples, and the second term computes the average binary cross-entropy loss for fake data samples.

## 3.4 Implementation details

During the training process, a series of preprocessing steps have been applied to enhance the performance and generalization of our proposed network for visual saliency detection.

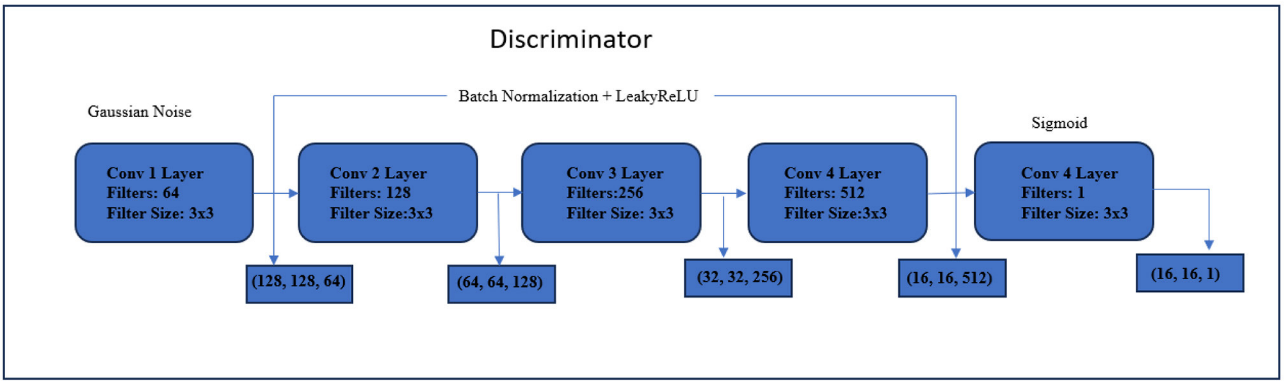

FIGURE 4
Discriminator module.

**Image preprocessing:** each input image is initially resized to a resolution of 256 × 256 pixels.

**No existing backbones:** our proposed network does not rely on any pre-existing backbone architectures. Instead, we design and implement an encoder-decoder architecture suitable for training the generator. A total of 8,470,850 trainable parameters are used for training the model.

**Data normalization:** all input data is normalized before training to ensure stable and consistent learning. Normalization is performed by subtracting the mean and scaling the images by the standard deviation. This normalization process helps the model converge more effectively and mitigates issues related to input distribution disparities. The training parmeters for cGAN are given in Table 1.

TABLE 1 Training parameters for cGAN.

| Number of epochs | 50 (Early stopping callback yielded 17.5 epochs) |
|---|---|
| Batch size | 32 |
| Optimizer | Adam |
| Learning rate | 0.001 for generator and 0.0001 for discriminator |
| Momentum of $\beta_1$ and $\beta_2$ | 0.5 and 0.999 |
| Loss function Margin + L2 Loss | Adversarial Loss, BCE Loss and L2 loss |

## 3.5 Technical details

The proposed architecture was implemented using Python 3 and TensorFlow framework. Training and testing were conducted on a single NVIDIA Tesla V100—SXM2 Graphical Processing Unit with 16 GB of memory. For initialization, the pre-trained enhanced U-Net model on the DUTS-TR dataset is used. The generator's weights were initialized with the pre-trained U-Net weights, while the discriminator was randomly initialized using a Gaussian distribution with a standard deviation of 0.002. During training, Adam optimization was employed with a learning rate of 1×10−3 for the generator and 1×10−4 for the discriminator. In the testing phase, the thoroughly trained generator was utilized to predict saliency maps. During the evaluation stage, we utilized a generator that is well-trained from our model to predict saliency maps. The generator, with its learned weights, was able to converge faster and produce accurate, reliable saliency maps, as depicted in Figure 5. To enhance the generalization ability of our model and prevent overfitting, we implemented early stopping with patience of 10. Early stopping is a technique that monitors the validation loss during training and terminates the training if the loss does not improve for a certain number of consecutive epochs. We utilized the Early Stopping callback from TensorFlow's Keras library, configuring it with patience of 10. Moreover, we set the "*restore best weights*" parameter to True, ensuring that the model's weights are restored to the best observed weights when the training is prematurely halted. This technique helps in preventing overfitting by terminating the process of training when the model's efficacy on the validation set reaches a plateau. It ensures that the model does not continue to learn from noise or irrelevant patterns in the data, thereby improving its ability to generalize to unseen examples. By leveraging the power of TensorFlow, the pre-trained U-Net model, and our carefully chosen training and optimization strategies, the proposed model is effective in the prediction of saliency maps.

## 3.6 Ablation study

In this section, we explore the individual contributions of each module. To determine the effectiveness of each module, we separately evaluate U-Net, our enhanced U-Net, and the fusion of enhanced U-Net and cGAN. These evaluations are conducted across three datasets, using maxF and MAE as our evaluation metrics. The study involved systematically removing or modifying specific model elements and evaluating their impact on performance. The following ablation experiments were performed and are given in Table 2:

**Baseline U-Net**: we established a baseline by training the original U-Net architecture without any modifications. This serves as a reference point for comparing the performance of subsequent variations.

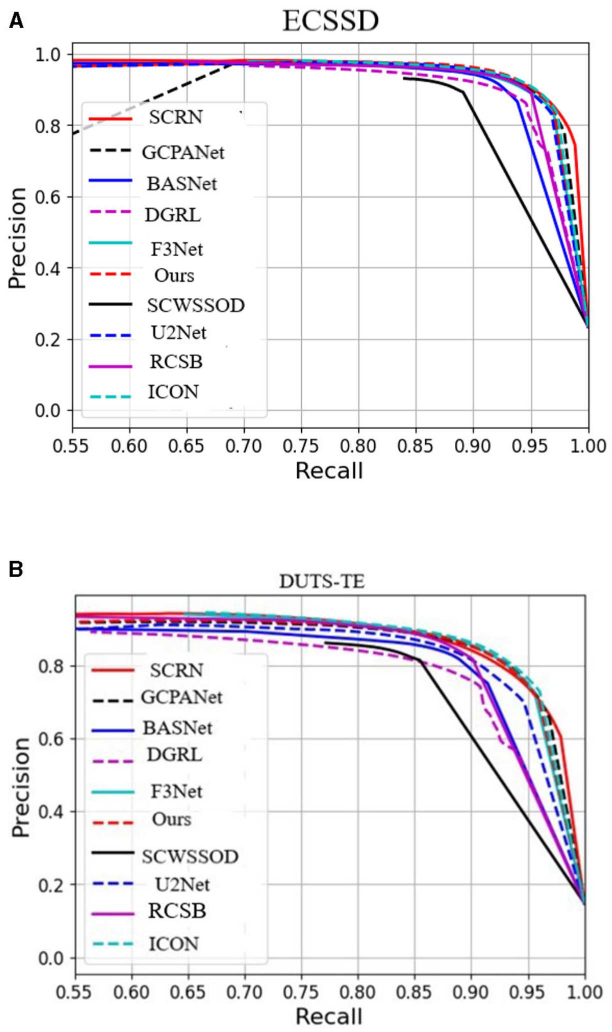

FIGURE 5
PR plot on ECSSD dataset is shown in **(A)** and PR plot on DUTS-TE dataset is shown in **(B)**.

**Spatial attention module**: to investigate the contribution of the spatial attention mechanism, we trained an enhanced U-Net model including the spatial attention module. This allows us to assess the influence of spatial attention on saliency detection accuracy.

**Dropout regularization**: we trained an enhanced U-Net model by excluding the dropout layers in the convolutional blocks. This experiment aims to assess the impact of dropout regularization on the model's generalization ability and resistance to overfitting, with dropout set to 0.2.

**Batch normalization**: an encoder-decoder module was trained by removing the batch normalization layers from both the encoder and decoder blocks. This experiment helps evaluate the effect of batch normalization on training stability and convergence.

We used the same training and evaluation protocols for each ablation experiment, including the loss function, optimizer, and evaluation metrics. Both models underwent training using the identical dataset and were subsequently assessed using established metrics such as accuracy, precision, recall, and F1-score. The outcomes of the ablation study yield valuable insights into how distinct elements and methods contribute to enhancing the saliency detection capabilities of our proposed architecture.

# 4 Experimental analysis

## 4.1 Datasets

To assess the performance of our proposed method, we conducted experiments on three popular benchmark datasets, including DUTS (Wang et al., 2017), ECSSD (Shi et al., 2015), and DUT-OMRON (Yang et al., 2013). The DUTS dataset contains two subsets namely DUTS-TR and DUTS-TE, comprise 10,553 images for training and 5,019 testing images respectively collected from different scenes. ECSSD is a complex scene dataset comprising 1,000 high-resolution images, while DUT-OMRON consists of 5,186 images derived from actual scenes.

## 4.2 Evaluation metrics

To comprehensively assess the quality compared to the ground truth, we utilize standard measures: sensitivity, specificity, F-score,

TABLE 2 Ablation study of individual module on ECSSD, DUTS-TE, DUT-OMRON in terms of MaxF (↑) and MAE (↓).

| Model | Dataset | | | | | |
|---|---|---|---|---|---|---|
| | ECSSD | | DUTS-TE | | DUT-OMRON | |
| | MaxF | MAE | MaxF | MAE | MaxF | MAE |
| UNet (Base model) | 0.7995 | 0.081 | 0.776 | 0.084 | 0.752 | 0.089 |
| Encoder-Decoder network (Enhanced U-Net with out cGAN) | 0.8135 | 0.068 | 0.821 | 0.062 | 0.826 | 0.075 |
| Encoder-Decoder network (Without added spatial attention) | 0.8014 | 0.072 | 0.801 | 0.069 | 0.793 | 0.082 |
| Fusion of cGAN<br>+ Encoder-Decoder + loss functions<br>+ (our proposed method) | 0.9375 | 0.0292 | 0.866 | 0.033 | 0.8192 | 0.043 |

TABLE 3 Quantitative comparison of the proposed method and ten state-of-the-art methods on three benchmark datasets is conducted in terms of MAE, $E_\xi$, $S_\alpha$, and $F_{\beta w}$.

| Dataset | Method | MAE↓ | $E_\xi$ ↑ | $S_\alpha$ ↑ | $F_{\beta w}$ ↑ |
|---|---|---|---|---|---|
| ECSSD | **Ours** | 0.0292 | 0.9459 | 0.9322 | 0.9182 |
| | DGRL | 0.0415 | 0.9342 | 0.9065 | 0.8783 |
| | GCPANet | 0.0335 | 0.9377 | 0.9264 | 0.8973 |
| | SCRN | 0.0351 | 0.9356 | 0.9269 | 0.9014 |
| | RCSB | 0.0342 | 0.944 | 0.9224 | 0.9171 |
| | BASNet | 0.0362 | 0.9384 | 0.9163 | 0.9066 |
| | F3Net | 0.033 | 0.9429 | 0.9245 | 0.9119 |
| | SCWSSOD | 0.0498 | 0.9228 | 0.8836 | 0.8819 |
| | ICON | 0.0313 | 0.9497 | 0.9292 | 0.913 |
| | U2Net | 0.0315 | 0.9422 | 0.927 | 0.9113 |
| DUTS-TE | **Ours** | 0.0331 | 0.9132 | 0.8878 | 0.828 |
| | DGRL | 0.0504 | 0.8847 | 0.8467 | 0.7486 |
| | GCPANet | 0.037 | 0.904 | 0.8897 | 0.8102 |
| | SCRN | 0.0383 | 0.897 | 0.8847 | 0.8014 |
| | RCSB | 0.0348 | 0.9141 | 0.8819 | 0.8429 |
| | BASNet | 0.0468 | 0.8917 | 0.8661 | 0.7997 |
| | F3Net | 0.0348 | 0.9153 | 0.889 | 0.8233 |
| | SCWSSOD | 0.0487 | 0.8976 | 0.8426 | 0.7964 |
| | ICON | 0.0366 | 0.9201 | 0.8891 | 0.8253 |
| | U2Net | 0.0437 | 0.8931 | 0.8733 | 0.8007 |
| DUT-OMRON | **Ours** | 0.0439 | 0.886 | 0.8674 | 0.7859 |
| | DGRL | 0.0632 | 0.8449 | 0.8097 | 0.6835 |
| | GCPANet | 0.0566 | 0.8468 | 0.8375 | 0.7194 |
| | SCRN | 0.56 | 0.8482 | 0.8366 | 0.7139 |
| | RCSB | 0.0492 | 0.8575 | 0.835 | 0.7506 |
| | BASNet | 0.0565 | 0.8649 | 0.8362 | 0.743 |
| | F3Net | 0.0526 | 0.861 | 0.8381 | 0.7326 |
| | SCWSSOD | 0.0602 | 0.8563 | 0.812 | 0.7252 |
| | ICON | 0.0569 | 0.875 | 0.8443 | 0.7431 |
| | U2Net | 0.0544 | 0.867 | 0.8466 | 0.7486 |

The symbols ↑ and ↓ are used to indicate that a higher score and a lower score, respectively, represent better results.

E-measure, S-measure, and MAE. Each evaluation parameter is described below.

**Sensitivity:** Sensitivity, also known as the true positive rate or recall, evaluates the model's capability to correctly recognize positive instances (foreground) from the ground truth. It measures the fraction of positive pixels correctly classified as positive.

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

**Specificity:** Specificity measures the model's ability to correctly identify negative instances (background) from the ground truth. It assesses the fraction of negative pixels correctly classified as negative.

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

**F-measure** (Fan et al., 2018a): The F-measure rates for the binarized saliency map are computed with a threshold range of [0,255] and is given by

$$F_\beta = \frac{\left(1 + \beta^2\right) Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{12}$$

**Mean Absolute Error(MAE)** (Perazzi et al., 2012): The Precision-Recall curve does not include the fraction of pixels correctly classified as non-salient. The presence of pixels mistakenly labeled as salient leads the saliency map to perform worse, even though it is smooth and has greater values allocated to salient pixels. Using Mean Absolute Error (MAE) as suggested by Perazzi et al. (2012), we can overcome the limitation of using precision and recall. We analyse the mean absolute error (MAE) between the continuous saliency map M and the binary ground truth GT for a more fair comparison that considers these factors. The lower the MAE score, the closer the model is to ground truth, and the better the performance.

$$MAE = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \left| \hat{I}_{ij} - I_{ij} \right| \tag{13}$$

Where, H denotes the height and W denotes the width in the size of groundtruth map, $\hat{I}_{ij}$ and $I_{ij}$ denote the predicted and groundtruth map respectively.

**E-measure** (Fan et al., 2018b): The Mean E-measure is a metric that quantifies the likeness between the predicted map and the actual ground-truth map. Its definition is as follows,

$$E_\xi = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \varphi\left(i, j\right) \tag{14}$$

where $\varphi(i,j)$ denote alignment matrix.

**S-measure** (Fan et al., 2017): The S-measure is a metric that quantifies structural similarity by combining the similarity and error measures of the area-conscious $S_r$ and object-conscious $S_0$ methods. A higher $S_\alpha$ value indicates superior algorithm performance. The computation formula is provided below:

$$S_\alpha = (1 - \alpha)S_r + \alpha S_0 \tag{15}$$

An open source implementations of PydenseCRF [1] and SOD Evaluation metrics [2] are adopted in this paper. [1], [2]

## 5 Results and discussion

The proposed enhanced U-Net model for salient object detection has been tested qualitatively and quantitatively against DGRL (Wang et al., 2018) (e) GCPANet (Chen et al., 2020) (f)

[1] https://github.com/lucasb-eyer/pydensecrf.git

[2] https://github.com/zyjwuyan/SOD_Evaluation_Metrics.git

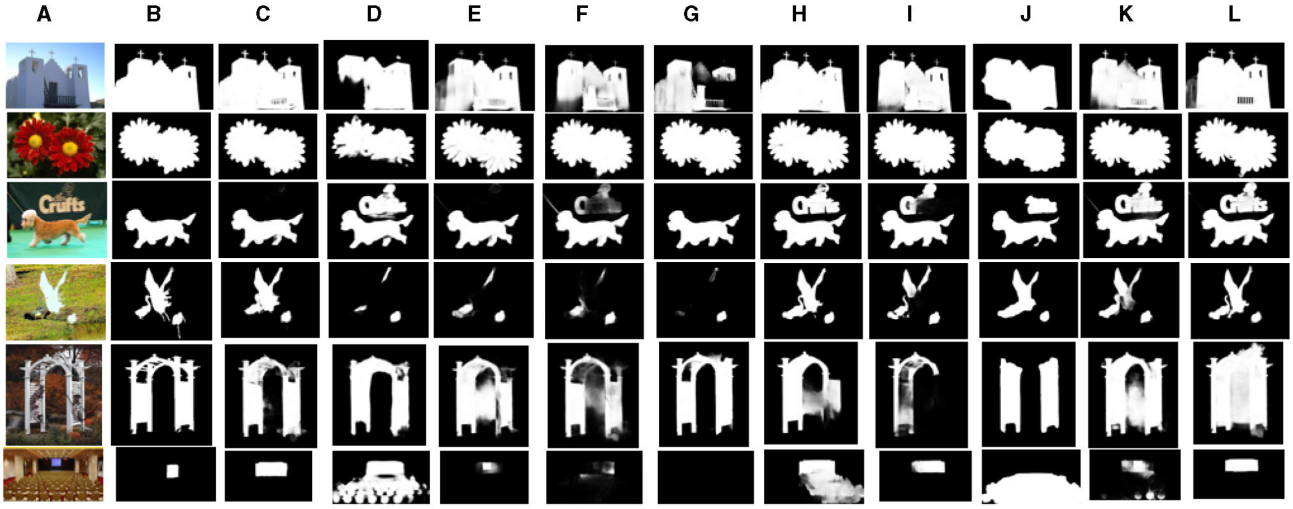

FIGURE 6
Qualitative analysis of our proposed approach with ten state-of-the-art methods. **(A)** Input image **(B)** Groundtruth **(C)** Ours **(D)** DGRL (Wang et al., 2018) **(E)** GCPANet (Chen et al., 2020) **(F)** SCRN (Wu et al., 2019) **(G)** RCSB (Ke and Tsubono, 2022) **(H)** BASNet (Qin et al., 2019) **(I)** F3Net (Jun Wei, 2020) **(J)** SCWSSOD (Yu et al., 2021) **(K)** ICON (Zhuge et al., 2022) **(L)** U2Net (Qin et al., 2020).

SCRN Wu et al. (2019) (g) RCSB (Ke and Tsubono, 2022) (h) BASNet (Qin et al., 2019) (i) F3Net (Jun Wei, 2020) (j) SCWSSOD (Yu et al., 2021) (k) ICON (Zhuge et al., 2022) (l) U2Net (Qin et al., 2020) methods. We conducted a fair comparison of each method within our environment, utilizing the publicly available source code provided by the authors for evaluation purposes. Figure 5 illustrates the qualitative results. The proposed method achieves good F-measure values even on complex image datasets like DUT-OMRON, outperforming state-of-the-art methods.We compare the performance of our method with other salient object detection approaches in terms of MAE, E-measure, S-measure and Max $F_\beta$ as shown in Table 3.

## 5.1 Qualitative results

To further emphasize the advantages of our proposed method, we present visual examples showcasing its efficacy in addressing various challenging scenarios. As depicted in Figure 6, our method demonstrates robust performance in handling fine-grained structures (2nd image), cluttered backgrounds (3rd and 4th images), complex structures (5th image), object concurrency (first row), and the presence of multiple salient objects. In comparison with state-of-the-art methods, the saliency maps generated by our approach exhibit enhanced completeness while maintains good accuracy. It is noteworthy that our method excels in managing background/foreground disturbances, as evidenced by the third row, and effectively captures relationships among multiple objects, as demonstrated in the second row. Most of the methods detect two objects instead of a single salient object, as shown in the third row and in the sixth row. However, in our method, the attention mechanism effectively highlights the salient object, with the prominent object focussed. The computational time for processing a single image using our proposed method is about 6 seconds. These results emphasize the effectiveness of our proposed method in aggregating features and integrating global context information.

## 5.2 Quantitative results

Table 3 presents the numerical outcomes from three conventional benchmark datasets. In this comparison, our approach is evaluated against 10 cutting-edge algorithms using metrics such as S-measure, weighted F-measure, E-measure, and MAE. The results clearly indicate that our model outperforms the other baseline methods.

In addition to the numerical comparisons in Table 3, Figures 6a, 6b, 7a, 7b, 8a, 8b, visualizes the performance of all compared methods using precision-recall and F-measure curves on three datasets. The dotted red line, representing our proposed method, consistently outperforms all others across most decision thresholds. This is likely due to the proposed method's effective use of complementary spatial information, which leads to more accurate localization, ultimately resulting in a superior precision-recall curve.

# 6 Conclusion

In this study, we present an efficient framework for SOD in images, effectively addressing the limitations of cGAN and U-Net architectures. The experimental analysis reveals that the proposed method performs well, with lower MAE values compared to other state-of-the-art approaches. Although the proposed SOD model outperformed all other baseline models, it required a substantial training time of approximately one hour. However, the inference time for most of the models is similar, taking around 6 seconds per image. Faster convergence is observed while training the generator with the introduction of enhanced U-Net model weights. With increased computing power, the proposed method could potentially yield faster results. Achieving high performance in low-contrast environments or with data where the distinction between object and background is unclear remains challenging. In future work, we aim to apply transfer learning. This approach can significantly

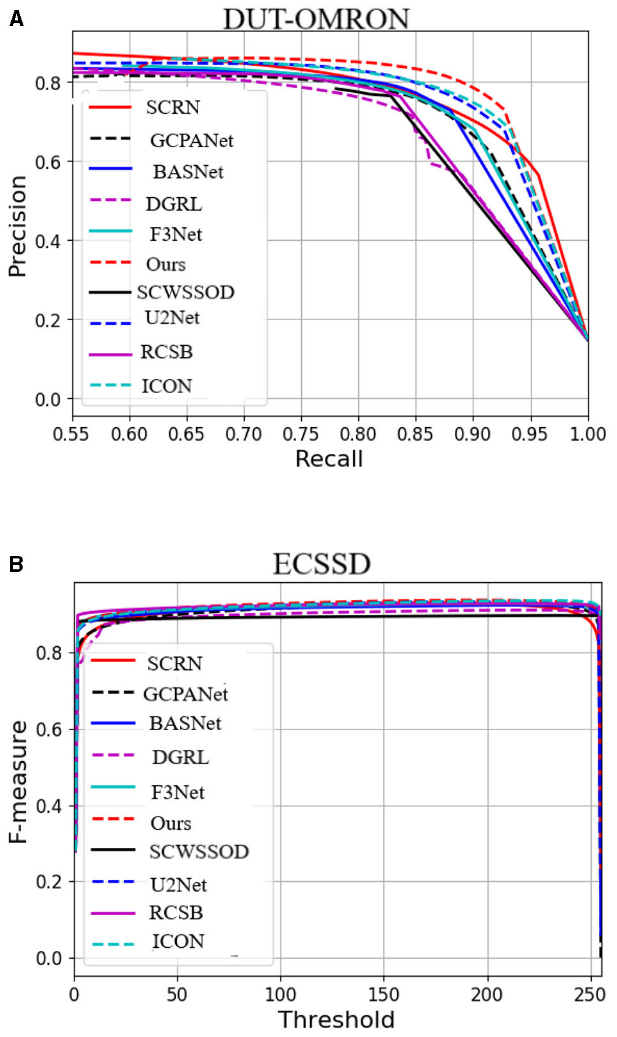

FIGURE 7
PR plot on DUR-OMRON dataset is shown in **A** and F-measure at different thresholds on ECSSD dataset is shown in **B**.

reduce the training time and computational resources required, as the model has already learned useful features from a large dataset. These improvements pave the way for enhanced visual understanding and interpretation in complex scenes, ultimately benefiting computer vision systems.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

## Author contributions

GD: Conceptualization, Data curation, Formal analysis, Methodology, Resources, Software, Writing – original draft, Writing – review & editing. RK: Methodology, Validation, Writing – review & editing.

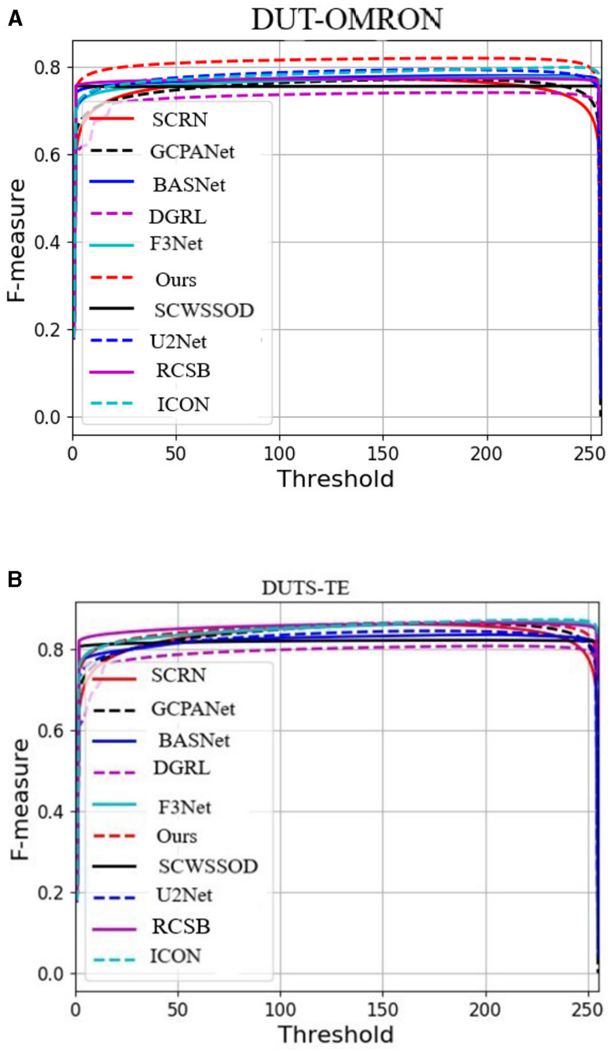

FIGURE 8
F-measure at different thresholds on DUT-OMRON dataset is as shown in **A** and F-measure at different thresholds on DUTS-TE dataset is as shown in **B**.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. We gratefully acknowledge the full financial support for this research work provided by SRM University AP Andhra Pradesh, India.

## Acknowledgments

We thank SRM University—AP for providing facilities in research lab for completing this research work.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

Aamir, M., Rahman, Z., Abro, W. A., Bhatti, U. A., Dayo, Z. A., and Ishfaq, M. (2023). A progressive approach to generic object detection: a two-stage framework for image recognition. *Comp. Mater. Continua* 75, 6351–6373. doi: 10.32604/cmc.2023.038173

Arjovsky, M., Chintala, S., and Bottou, L. (2017). "Wasserstein generative adversarial networks," in *International Conference on Machine Learning* (New York: PMLR), 214–223.

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 2481–2495. doi: 10.1109/TPAMI.2016.2644615

Borji, A., and Itti, L. (2012). State-of-the-art in visual attention modeling. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 185–207. doi: 10.1109/TPAMI.2012.89

Buric, M., Pobar, M., and Ivasic-Kos, M. (2018). "Ball detection using yolo and mask R-CNN," in 2018 *International Conference on Computational Science and Computational Intelligence (CSCI)* (Las Vegas, NV: IEEE), 319–323.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv [Preprint]*. arXiv:1412.7062. doi: 10.48550/arXiv.1412.7062

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2017a). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 834–848. doi: 10.1109/TPAMI.2017.2699184

Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. (2017b). Rethinking atrous convolution for semantic image segmentation. *arXiv [Preprint]*. arXiv:1706.05587. doi: 10.48550/arXiv.1706.05587

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Cham: Springer), 801–818. doi: 10.1007/978-3-030-01249-6

Chen, Q., Liu, T., Shang, Y., Shao, Z., and Ding, H. (2019). Salient object detection: Integrate salient features in the deep learning framework. *IEEE Access* 7, 152483–152492. doi: 10.1109/ACCESS.2019.2948062

Chen, Z., Xu, Q., Cong, R., and Huang, Q. (2020). Global context-aware progressive aggregation network for salient object detection. *Proc. AAAI Conf. Artif. Intellig.*34, 10599–10606. doi: 10.1609/aaai.v34i07.6633

Cheng, M.-M., Mitra, N. J., Huang, X., Torr, P. H., and Hu, S.-M. (2014). Global contrast based salient region detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 569–582. doi: 10.1109/TPAMI.2014.2345401

Fan, D.-P., Cheng, M.-M., Liu, J.-J., Gao, S.-H., Hou, Q., and Borji, A. (2018a). "Salient objects in clutter: Bringing salient object detection to the foreground," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Cham: Springer), 186–202.

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

Fan, D.-P., Gong, C., Cao, Y., Ren, B., Cheng, M.-M., and Borji, A. (2018b). Enhanced-alignment measure for binary foreground map evaluation. *arXiv [Preprint]*. arXiv:1805.10421. doi: 10.24963/ijcai.2018/97

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE), 580–587. doi: 10.1109/CVPR.2014.81

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 27.

Guan, Y., Aamir, M., Hu, Z., Dayo, Z. A., Rahman, Z., Abro, W. A., et al. (2021). An object detection framework based on deep features and high-quality object locations. *Traitement du Signal* 38:19. doi: 10.18280/ts.380319

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). "Improved training of Wasserstein GANs," in *Advances in Neural Information Processing Systems*, 30.

Han, L., Li, X., and Dong, Y. (2019). Convolutional edge constraint-based U-Net for salient object detection. *IEEE Access* 7, 48890–48900. doi: 10.1109/ACCESS.2019.2910572

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 770–778.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 1125–1134.

Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., and Bengio, Y. (2017). "The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (IEEE), 11–19. doi: 10.1109/CVPRW.2017.156

Ji, Y., Zhang, H., and Wu, Q. J. (2018a). Saliency detection via conditional adversarial image-to-image network. *Neurocomputing* 316, 357–368. doi: 10.1016/j.neucom.2018.08.013

Ji, Y., Zhang, H., and Wu, Q. J. (2018b). Salient object detection via multi-scale attention cnn. *Neurocomputing* 322:130–140. doi: 10.1016/j.neucom.2018.09.061

Jiang, B., Zhang, L., Lu, H., Yang, C., and Yang, M.-H. (2013a). "Saliency detection via absorbing markov chain," in *Proceedings of the IEEE International Conference on Computer Vision* (Sydney, NSW: IEEE), 1665–1672. doi: 10.1109/ICCV.2013.209

Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N., and Li, S. (2013b). "Salient object detection: A discriminative regional feature integration approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Portland, OR: IEEE), 2083–2090. doi: 10.1109/CVPR.2013.271

Jun, W., and Shuhui Wang, Q. H. (2020). "F3net: Fusion, feedback and focus for salient object detection," in *AAAI Conference on Artificial Intelligence (AAAI)* (PKP Publishing Services Network). doi: 10.1609/aaai.v34i07.6916

Ke, Y. Y., and Tsubono, T. (2022). "Recursive contour-saliency blending network for accurate salient object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (Waikoloa, HI: IEEE), 2940–2950. doi: 10.1109/WACV51458.2022.00143

Koch, C., and Ullman, S. (1987). "Shifts in selective visual attention: towards the underlying neural circuitry," in *Matters of Intelligence: Conceptual Structures in Cognitive Neuroscience* (Cham: Springer), 115–141.

Krähenbühl, P., and Koltun, V. (2011). "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in Neural Information Processing Systems* (Curran Associates, Inc), 24.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* (Curran Associates, Inc.), 25.

Li, Y., Hou, X., Koch, C., Rehg, J. M., and Yuille, A. L. (2014). "The secrets of salient object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Columbus, OH: IEEE), 280–287.

Lian, X., Pang, Y., Han, J., and Pan, J. (2021). Cascaded hierarchical atrous spatial pyramid pooling module for semantic segmentation. *Pattern Recognit.* 110:107622. doi: 10.1016/j.patcog.2020.107622

Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., et al. (2010). Learning to detect a salient object. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 353–367. doi: 10.1109/TPAMI.2010.70

Mino, A., and Spanakis, G. (2018). "Logan: Generating logos with a generative adversarial neural network conditioned on color," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* (Orlando, FL: IEEE), 965–970.

Pang, Y., Zhao, X., Zhang, L., and Lu, H. (2020). "Multi-scale interactive network for salient object detection," in *Proceedings of the IEEE/CVF Conference On Computer Vision and Pattern Recognition* (Seattle, WA: IEEE), 9413–9422.

Perazzi, F., Krähenbühl, P., Pritch, Y., and Hornung, A. (2012). "Saliency filters: contrast based filtering for salient region detection," in *2012 IEEE Conference on Computer Vision and Pattern Recognition* (Providence, RI: IEEE), 733–740.

Qi, M., and Wang, Y. (2016). "Deep-CSSR: Scene classification using category-specific salient region with deep features," in *2016 IEEE International Conference on Image Processing (ICIP)* (Phoenix, AZ: IEEE), 1047–1051.

Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O. R., and Jagersand, M. (2020). U2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognit*. 106:107404. doi: 10.1016/j.patcog.2020.107404

Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., and Jagersand, M. (2019). "BASNet: Boundary-aware salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA: IEEE), 7479–7489.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). "Faster R-CNN: towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems* (Curran Associates, Inc.), 28.

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference, Munich, Germany* (Cham: Springer), 234–241.

Shi, J., Yan, Q., Xu, L., and Jia, J. (2015). Hierarchical image saliency detection on extended cssd. *IEEE Trans. Pattern Anal. Mach. Intell.* 38, 717–729. doi: 10.1109/TPAMI.2015.2465960

Simonyan, K., and Zisserman, A. (2015). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. IEEE.

Singh, N. (2020). Saliency threshold: a novel saliency detection model using ising's theory on ferromagnetism (stif). *Multimedia Syst.* 26, 397–411. doi: 10.1007/s00530-020-00650-z

Sun, Y., Zhao, M., Hu, K., and Fan, S. (2022). Visual saliency prediction using multi-scale attention gated network. *Multimedia Syst.* 28, 131–139. doi: 10.1007/s00530-021-00796-4

Treisman, A. M., and Gelade, G. (1980). A feature-integration theory of attention. *Cogn. Psychol.* 12, 97–136. doi: 10.1016/0010-0285(80)90005-5

Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., et al. (2017). "Learning to detect salient objects with image-level supervision," in *CVPR* (IEEE).

Wang, T., Zhang, L., Wang, S., Lu, H., Yang, G., Ruan, X., et al. (2018). "Detect globally, refine locally: A novel approach to saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE),3127–3135.

Wang, Y., Wang, R., Fan, X., Wang, T., and He, X. (2023). "Pixels, regions, and objects: Multiple enhancement for salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Vancouver, BC: IEEE), 10031–10040.

Wei, Y., Wen, F., Zhu, W., and Sun, J. (2012). "Geodesic saliency using background priors," in *Computer Vision-ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III 12* (Cham: Springer), 29–42. 3.

Wu, Z., Su, L., and Huang, Q. (2019). "Stacked cross refinement network for edge-aware salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Seoul: IEEE), 7264–7273.

Xie, S., and Tu, Z. (2015). "Holistically-nested edge detection," in *Proceedings of the IEEE International Conference on Computer Vision* (IEEE), 1395–1403.

Yang, C., Zhang, L., Lu, H. R. X., and Yang, M.-H. (2013). "Saliency detection via graph-based manifold ranking," in *Computer Vision and Pattern Recognition (CVPR)* (Portland, OR: IEEE), 3166–3173. doi: 10.1109/CVPR.2013.407

Yu, S., Zhang, B., Xiao, J., and Lim, E. G. (2021). Structure-consistent weakly supervised salient object detection with local saliency coherence. *Proc. AAAI Conf. Artif. Intellig.* 35, 3234–3242. doi: 10.1609/aaai.v35i4.16434

Zhang, P., Liu, W., Lu, H., and Shen, C. (2019). Salient object detection with lossless feature reflection and weighted structural loss. *IEEE Trans. Image Proc.* 28, 3048–3060. doi: 10.1109/TIP.2019.2893535

Zhang, P., Wang, D., Lu, H., Wang, H., and Ruan, X. (2017). "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice: IEEE), 202–211.

Zhao, H., Qi, X., Shen, X., Shi, J., and Jia, J. (2018). "ICNet for real-time semantic segmentation on high-resolution images," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Springer International Publishing), 405–420.5

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017). "Pyramid scene parsing network," in *Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition* (Venice: IEEE), 2881–2890. doi: 10.1109/CVPR.2017.660

Zhao, J.-X., Liu, J.-J., Fan, D.-P., Cao, Y., Yang, J., and Cheng, M.-M. (2019). "Egnet: Edge guidance network for salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Seoul: IEEE), 8779–8788.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., et al. (2015). "Conditional random fields as recurrent neural networks," in *Proceedings of the IEEE International Conference on Computer Vision* (Washington, DC: IEEE), 1529–1537.

Zhu, W., Liang, S., Wei, Y., and Sun, J. (2014). "Saliency optimization from robust background detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Columbus, OH: IEEE), 2814–2821.

Zhuge, M., Fan, D.-P., Liu, N., Zhang, D., Xu, D., and Shao, L. (2022). Salient object detection via integrity learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 3738–3752. doi: 10.1109/TPAMI.2022.3179526

Zünd, F., Pritch, Y., Sorkine-Hornung, A., Mangold, S., and Gross, T. (2013). "Content-aware compression using saliency-driven image retargeting," in *2013 IEEE International Conference on Image Processing* (Melbourne, VIC: IEEE), 1845–1849.