Attenuative Unified U-Net with Conditional PatchMatch Algorithm for Edge Preserving Salient Object Detection

GAURAV GUPTA, ARUN K. SUNANIYA
National Institute of Technology, Silchar, Assam-788010, INDIA

Abstract: The rise of Deep Neural Networks (DNNs) has significantly boosted salient object detection in computer vision tasks. However, downsampling methods like striding and pooling often introduce bluish artifacts near edges, hindering detection accuracy. To address this, the "Attenuative Unified U-Net with Conditional PatchMatch Algorithm" is proposed. The method employs Gaussian smoothing for noise reduction and depth refinement in preprocessing. In existing methods, the Edge-preserving salient object detection struggles with cluttered backgrounds and connected objects, like a bird on a rock, making it difficult to accurately distinguish edges and salient regions due to shared boundaries. Hence, an Attenuative Unified Backpropagated Entropy U-Net is proposed, which integrates an Attention Mechanism that enhance feature map spatial weight assignment, emphasizing features or regions that are judged most relevant for edge and salient region detection. Then, the Cascaded Laplace pyramid is incorporated into the Unified U-Net's design for multi-scale information processing, capturing contextual details effectively. The Backpropagated Entropy Loss Function is then created to ensure accurate and confident fused image output. The depth map, edge map, and salient area map are combined in the U-Net's final output layer to create an overall fused image. Next, to addresses depth discontinuity problem a noteworthy novel idea the Conditional PatchMatch Random Field Algorithm (CPMRF), which combines PatchMatch efficiency and CRFs' contextual modeling. PatchMatch iteratively propagates matches from neighboring patches to find similar patches in the remaining portion of the image and Conditional Random Field (CRF) is the next step, which improves the PatchMatch results by taking into account the connections between the matched patches and their neighboring areas. Hence, the experimental outcomes of the proposed model effectively show the improved detection of edge in salient object detection with better accuracy, precision, recall, MaxF, and F1 score with minimized MAE.

Keywords: Salient Object Detection, U-Net, PatchMatch, Gaussian smoothening, Laplace pyramid, Conditional Random Field

Received: April 2, 2024. Revised: September 7, 2024. Accepted: September 24, 2024. Published: October 17, 2024.

1. Introduction

The two main categories of saliency detection [1-2] are salient object recognition and eye fixation prediction. The former seeks to detect the most salient elements in an image, while the latter predicts the location of the human eye's next fixation. The rapid development of deep learning in recent years has led to considerable advancements in saliency detection. Finding visually interesting or salient objects in a scene is one of the main tasks in the science of computer vision. It can be used for many different visual tasks, such as segmentation [3], video compression [4], person re-identification [5], and recognition of human activities. To enhance the effectiveness of salient object recognition, a multitude of RGB saliency models, including traditional methods and methods based on convolutional neural networks (CNNs), have been created recently. Nevertheless, the current RGB saliency models doesn't produce sufficient saliency maps in challenging situations like similar appearances of salient objects and background, cluttered backdrops, diverse salient objects, and so forth [6-7].

To take advantage of depth signals, the first difficulty is to effectively extract features from depth maps. The early standard RGBD models [8] were limited in their ability to represent high-level data, therefore they relied mostly on domain expertise to build hand-crafted features. As proof of CNNs' supremacy in the interim, numerous CNN-based RGBD models are offered to automatically learn deep depth features from the depth maps by employing the widely utilized feature extraction networks. The majority of these RGBD saliency models use pre-trained models VGG-16 or ResNet to extract their deep depth features; these models are trained from a large-scale RGB image dataset called ImageNet [10–11]. However, some of these models [9] even use a Siamese architecture, sharing the parameters in the RGB and depth branches. There is undoubtedly a big difference between the intrinsic properties of RGB and depth information because

depth maps don't have as much color and texture information. As a result, the extraction of depth features may become unclear whether the Siamese architecture or the pre-trained models are fine-tuned.

Next, it is crucial to focus more on the second issue with RGBD saliency detection [12–13], which is the effective fusion of RGB and depth features. Certainly, a great deal of work has been done to look into the two modalities' fusion mechanism. For example, early efforts attempted to use linear summing to merge RGB and depth data. For the CNN-based RGBD saliency models, they design symmetry architecture to establish relationships between RGB properties and depth features. A detailed analysis shows that the fusion method takes into account the depth RGB and deep RGB data in an equal manner. However, as previously said, there are significant differences in the underlying attribution of the two modes. Consequently, contemporary efforts distinguish between the two modalities [14–15], that is, concentrating more on RGB information or depth information, in an attempt to blend RGB characteristics with depth data. RGBD saliency detection performance is significantly enhanced by doing this. However, while handling some challenging scenes, the state-of-the-art models' performance will also deteriorate. Therefore, new methods for edge-preserving salient object detection from RGBD images under depth detection are required. The following are the paper's main contribution:

- To detect the edge map and depth map with salient region and to fuse them together a novel Attenuative Unified U-Net with Conditional PatchMatch Algorithm is used.
- To address the depth discontinuity issue and improves the output with accurate salient object prediction a novel Conditional PatchMatch Random Field Algorithm is used.

The proposed work is divided into five sections, the first of which is an introduction and the second is a review of the literature. Section 3 encapsulates the proposed method. It also discusses the various research methods that have been used. Performance and comparative analysis are covered in Section 4. Section 5 delves deeply into the work's conclusion.

2. Literature Survey

Zhou et al. [16] presented a simple yet effective technique called Hierarchical and Interactive Refinement Network (HIRN) for maintaining edge structures while identifying important objects. For high-level and low-level feature maps, a dual-path and multi-stage network structure was constructed to estimate the salient areas and edges. As a result, projected regions improved by strengthening weak responses at edges, while anticipated edges gained greater significance by decreasing false positives in the background. After the salient maps of edges and regions at the output layers were obtained, an edge-guided inference technique was applied to further filter the generated regions along the expected edges. This network's complex architecture made it computationally expensive, which limited its real-time applications.

Tu et al. [17] proposed a unique Edge-guided Non-local FCN (ENFNet) to achieve edge-guided feature learning for precise salient object recognition while preserving the distinct edge structure of salient objects. In particular, to integrate non-local characteristics for efficient feature representations in FCN, hierarchical global and local information was retrieved. To further maintain appropriate borders of salient objects, a guidance block was proposed to insert edge prior information into hierarchical feature maps. The guidance block performs both feature-wise and spatial-wise adjustment for successful edge embedding. To aid in improving the performance of salient item detection, future research will look into alternate preceding information or knowledge, such as semantic priors or thermal infrared data. It will also extend this approach to salient object detection in films.

Wen et al. [18] implemented Dynamic Selective Network (DSNet), which was a RGB-D saliency model, to perform salient object detection. To locate significant things roughly, a cross-modal global context module (CGCM) was first deployed to gather high-level semantic information. Next, by using gated and pooling based selection, the multi-level and multi-scale information was further optimized. In order to do this, a dynamic selective module (DSM) was created in order to dynamically mine the complimentary cross-modal information between depth maps and RGB images. It was employed to solve the problem of using RGB and depth information to detect salient objects in images. However, it was not effective when there were large fluctuations in the lighting or when there are occlusions in the scene.

Sun et al. [19] introduced an Average and Max-Pool Network (AMPNet) that integrates the multilevel complimentary contextual features by using the average- and max-pool modules. To increase the precision of identifying salient items, two top-down feedback channels were added, and their top-level semantic guidance information was completely exploited. Ultimately, the network's performance was enhanced across various datasets through the

utilization of the Feature Fusion Module and Deep Supervision Mechanism. To increase the completeness of items spotted, the network generates complementary contextual characteristics. Nevertheless, the spatial information is not entirely preserved by the pooling algorithms, leading to a loss of specific details about the prominent objects.

Li et al. [20] presented an edge information-guided hierarchical feature fusion network (HFFNet). This network effectively kept unambiguous edge information and correct semantic information while fusing features in a hierarchical fashion. A one-to-one hierarchical supervision technique was modified to oversee the creation of high-level and low-level information to improve retention of information at various levels. Ultimately, the creation of the saliency map was guided by low-level edge information, and saliency regions were identified by the edge guidance fusion. It was applied to enhance the precision of salient object recognition by efficiently integrating characteristics at various levels of abstraction. However, in difficult situations, the hierarchical feature fusion method was unable to fully capture all the minute details of the salient object in the image.

Wu et al. [21] introduced a Dynamic Scale Routing for Salient Object Detection. Initially, the dynamic pyramid convolution (DPConv), which vigorously chose the optimal kernel sizes, was employed for the encoder design in place of the vanilla convolution with fixed kernel sizes. Second, to best suit the DPConv-based encoder, a self-adaptive bidirectional decoder design was offered. The primary significance was its ability to route across feature scales and their dynamic collection, which made the inference process scale-aware. However, it has trouble in correctly identifying salient objects in complex scenes with cluttered backgrounds.

Kang et al. [22] presented a progressive and multi-task learning scheme, to excerpt the object context by only operating the learning scheme without altering the network architecture. A technique to gradually develop the decoder in the train phase was the progressive learning scheme. A multi-task learning (MTL) strategy that concurrently processed the contour and object saliency map in a single network was trained using the MTL approach via an auxiliary branch that learned contours in an edge-preserved manner. Other techniques for manipulating convolution blocks were integrated with the learning scheme. However, training takes a large amount of time and computer resources due to the progressive and multi-task learning approach.

Lad et al. [23] proposed a Boundary Preserved Salient Object Detection by means of Guided Filter based Spatial Domain Analysis and Hybridization Approach of Transformation. To create the final saliency map, it entailed super pixel segmentation, wavelet-based and learning-based saliency map production, and guided filter fusion of the two. It offers computational efficiency and preserved object boundaries, making it useful for tasks where accurate object detection with preserved boundaries were important. However, it failed to detect specific salient regions when both the learning-based and transformation-based saliency maps were incapable to notice them simultaneously.

Xu et al. [24] implemented a Multi-Stream Attention-Aware Graph Convolution Network for Video Salient Object Detection. The spatiotemporal graph creation, smoothness-aware regularization, attention-aware module, and graph convolution operation were the steps in the procedure. It aimed to address the challenge of preserving salient object boundaries and enhancing learning ability in video salient object detection. However, there was a need to improve the model's capacity to learn well and maintain salient object boundaries.

Liu et al. [25] introduced a Stereoscopically Attentive Multiscale (SAM) module, which adopted a stereoscopic attention mechanism to adaptively fuse the features of various scales. SAMNet used a lightweight architecture, multi-scale feature extraction, and stereoscopic information to accurately and efficiently recognize salient objects in stereoscopic pictures. It was appropriate for real-world applications with computing limitations due to its capacity to efficiently utilize stereoscopic information, extract multi-scale characteristics, and carry out accurate salient object recognition in a lightweight manner. However, its performance is subpar when it comes to conventional 2D images.

Zhang et al. [26] introduced an augmenting feedforward neural networks by using the multistage refinement mechanism. The majority of the intricate structures were absent from the coarse prediction map created in the first stage using a master net. The subsequent phases involved the refining of the previous saliency maps in a stage-by-stage manner, made possible by the refinement net's layer-wise recurrent connections to the master net and its ability to gradually incorporate local context information across stages. Moreover, the channel attention module and pyramid pooling module were used to combine global contexts depending on several regions. It was employed to efficiently recognize and draw attention to the most significant or salient objects in an image. However, the accuracy of the salient item recognition is impacted when dealing

with objects whose colors or textures are comparable to the background.

Zhou et al. [27] proposed end-to-end salient object detection in stereoscopic 3D images using Deep Convolutional Residual Autoencoder (DCRA). Multiple feature map fusion modules were built to examine the intricate linkages and take advantage of the complementarity among RGB and depth information. A convolutional residual module was useful to both the decoder and the encoder. Ultimately, a supervision pyramid depend on boundary loss and background previous loss was used to effectively optimize DCRA parameters. It was employed to precisely locate and extract significant sections or items from 3D images. However, it finds it difficult to accurately represent the depth information present in the 3D images.

Zhou et al. [28] introduced an Edge-aware Multi-Level Interactive Network, to identify the flaws from the strip steel surface. First, create a new connection where features from nearby encoder levels were transported to the similar stage of the decoder, excluding the skip connection, which merged the same stage of the encoder and decoder. Second, add the edge extraction branch after every decoder block to provide clearly defined bounds for expected outcomes. Edge-awareness and multi-level interaction were used in this strategy to efficiently identify and highlight regions of interest. However, it has trouble identifying extremely minute or intricate flaws in the surface of the strip steel.

Li et al. [29] introduced a lightweight network for ORSI-SOD depend on edge alignment and semantic matching, termed SeaNet. Semantic kernels were created by first compressing the high-level characteristics. Then, using dynamic convolution operations in DSMM, semantic kernels were employed to stimulate salient object locations in two groups of high-level characteristics. In the meantime, ESAM employed L2 loss to self-align the cross-scale edge data that was taken from two sets of low-level features and utilized it for detail augmentation. Lastly, the decoder infered prominent objects depend on the precise placements and fine details found in the outcomes of the two modules, starting with the highest-level features. Using edge alignment and semantic matching, this technique quickly and accurately identified significant items. However, it struggled to correctly identify the most salient object in photos with several overlapping objects or complex textures.

Wang et al. [30] proposed a deep subregion network (DSR-Net) equipped with a sequence of subregion dilated blocks (SRDB) by aggregating multi-scale salient context information of multiple sub-regions. A parallel ASPP module was created to enhance feature maps at each sub-region after SRDB divided the input feature map at various layers of a convolutional neural network (CNN) into distinct sub-regions. It analyzed sub-regions of an image using deep learning algorithms to identify the most important or visually striking sections. However, it viewed the salient objects with several non-salient interior holes as salient objects.

According to the talks above, the network's complicated structure made it computationally expensive for [16], which restrict real-time applications. Future studies on [17] investigate how to improve the effectiveness of salient item detection and expand the technique to identify items in films by utilizing semantic priors or thermal infrared data. [18] is not effective when there were large fluctuations in the lighting or when there were occlusions in the scene. As a first step toward [19], the spatial information is not entirely preserved by the pooling algorithms, leading to a loss of specific details about the prominent objects. The hierarchical feature fusion method was unable to fully capture all the minute details of the salient object in the image for [20]. [21], has trouble in correctly identifying salient objects in complex scenes with cluttered backgrounds. In [22], the training took a large amount of time and computer resources due to the progressive and multi-task learning approach. [23] fails to detect specific salient regions when both the transformation-based and learning-based saliency maps were unable to detect them simultaneously. In [24], there was a need to improve the model's capacity to learn well and maintain salient object boundaries. In [25], its performance was subpar when it comes to conventional 2D images and in [26] the accuracy of the salient item recognition was impacted when dealing with objects whose colors or textures were comparable to the background. [27] found it difficult to accurately represent the depth information present in the 3D images and [28] has trouble identifying extremely minute or intricate flaws in the surface of the strip steel. [29] struggled to correctly identify the most salient object in photos with several overlapping objects or complex textures. [30] viewed the salient objects with several non-salient interior holes as salient objects. Hence there is a need to develop a novel model to tackle these problems in the future.

3. Attenuative Unified U-Net with Conditional Patchmatch Algorithm for Edge Preserving Salient Object Detection

Salient object detection, which is usually taken as an important preprocessing procedure in various computer vision tasks, has undergone a very rapid development with the blooming of Deep Neural Network (DNN). However, the results of down sampling techniques like pooling and striding are invariably bluish near the margins, which has significantly impaired the performance of salient object detection. Hence, a novel "**Attenuative Unified U-Net with Conditional PatchMatch Algorithm for Edge Preserving Salient Object Detection**" has been developed to effectively addresses the aforementioned problems with edge-preserving salient object recognition. The current models for edge-preserving salient object detection have limitations when it comes to identifying the edge and salient regions from a cluttered background. This is because the scenario involves multiple connecting objects, such as a bird perched on a rock, and it is challenging to distinguish the edge from various connected objects due to shared boundaries. The low-level and high-level feature maps were alone not feasible for identifying the edge and salient region, and in situations with several connecting items, these feature maps cause discontinuities at the boundaries of connected objects, which results in fragmented edge maps. It is difficult to capture well-defined boundaries in more complicated objects with irregular shapes using low-level characteristics alone. Therefore, Attenuative Unified Backpropagated Entropy U-Net, a novel elucidate network, is presented. It incorporates an Attention Mechanism that learns to assign weights to various spatial locations or channels of the feature maps, emphasizing features or regions that are judged most relevant for edge and salient region detection. The Cascaded Laplace pyramid, which combines the series of Laplacian pyramids to decompose an image into a series of band-pass filtered versions with each level containing details at a specific scale, is incorporated into the Unified U-Net's design to handle multiple types of information (depth, edge, and salient regions) simultaneously. This enables the network to process the image at multiple scales, which is essential for capturing features at different sizes, thereby effectively captures the contextual information and depth features from the edge and salient region at different scales and levels of complexity. Then the Backpropagated Entropy Loss Function is developed to encourage the network to provide an overall fused image output that is both accurate and confident, by defining a weighting factor for each map (depth, edge and salient region) depending on its importance in varied parts of image. The weighted total of the individual backpropagated entropy losses is used to calculate the overall entropy loss. Backpropagation is used to update the model's weights during training in order to minimize the backpropagated entropy loss. The depth map, edge map, and salient area map are combined in the U-Net's final output layer to create an overall fused image. Moreover, the current cross-modal network is unable to handle the discontinuity problem because the edges found in the RGB color channels do not perfectly align with the depth discontinuities. This results in differences between the edge map and the depth map, making it difficult to accurately detect edges and fuse them with the salient region for improved salient object detection. Hence, a novel Conditional PatchMatch Random Field Algorithm for solving the depth discontinuity problem is presented. This algorithm combines the advantages of contextual modeling with the effectiveness of Patch Match with the ability to search for similar patches throughout the image by repeatedly propagating matches from nearby patches. The relationships between the matched patches and their surrounding regions are taken into account by Conditional Random Field, which improves the PatchMatch outcomes. Thus, by offering superior completions or modifications in depth map reconstruction, the CPRMF addresses the depth discontinuity issue and improves the U-Net output with accurate salient object prediction.

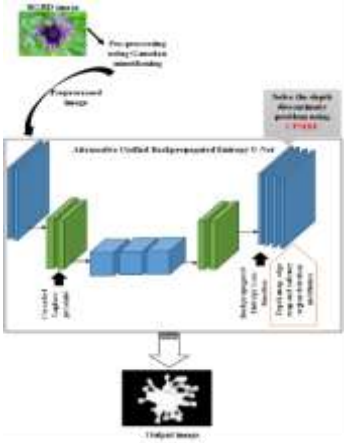


Fig.1. Block diagram of the proposed model

Fig.1 illustrates the block diagram of the proposed model. Salient object detection uses RGB-D picture data as the input. The detection process starts with preprocessing, where noise is reduced and the depth value is refined using Gaussian smoothening. Then,

the processed image is send to predict the depth and edge maps using the Attenuative Unified Backpropagated Entropy U-Net. The Cascaded Laplace pyramid, which combines the sequence of Laplacian pyramids to deconstruct an image into a succession of band-pass filtered variants, is incorporated into the Unified U-Net. Then, the Backpropagated Entropy Loss Function is defined to encourage the network to produce an overall fused image output that is both accurate and confident. Finally, the CPMRF algorithm is incorporated within U-Net, which solves the depth discontinuity problem by combining the benefits of Patch Match efficiency and CRFs' contextual modeling.

3.1. RGB-D Salient Object Detection

With a small modification, RGB-D salient object identification is directly performed using this structure. In contrast to RGB image salient object detection, RGB-D image salient object detection requires an additional input (depth). Consequently, use a backbone network to add depth feature inputs.

3.2. Pre-processing

Preprocessing is the first stage of salient object recognition, where Gaussian smoothening is used to lower noise and improve depth values. Let's represent the Gaussian smoothing process mathematically:

Given an input image I, the output of the Gaussian smoothing operation $I_{smoothed}$ is computed as per equation (1):

$$I_{smoothed}(x,y) = \sum_{i=-k}^{k} \sum_{j=-k}^{k} G(i,j,\sigma) \cdot I(x+i, y+j) \quad (1)$$

where, $I_{smoothed}(x,y)$ is the pixel value at position (x,y) in the smoothed image, $I(x+i,y+j)$ is the position's pixel value at $(x+i,y+j)$ in the original image, $G(i,j,\sigma)$ is the 2D Gaussian kernel having standard deviation σ, and k is the extent of the filter (kernel size). The kernel size determines the range of pixels to consider for the smoothing.

The 2D Gaussian kernel $G(i,j,\sigma)$ is defined in equation (2):

$$G(i,j,\sigma) = \frac{1}{2\pi\sigma^2} \cdot e^{-(i^2+j^2)/(2\sigma^2)} \quad (2)$$

Here, σ controls the spread of the Gaussian distribution, and a larger σ results in a wider smoothing effect. Thus, the convolution of the input image with a 2D Gaussian kernel produces a smoothed version of the image using the Gaussian smoothing process, which also helps to improve the quality of depth measurements and lower noise.

3.3. Attenuative Unified U-Net with Conditional PatchMatch Algorithm

The preprocessed image is given as the input to U-Net to detect the salient region with edge and depth features effectively. Biomedical picture segmentation is the focus of U-Net, a CNN architecture. Salient object detection and other semantic segmentation tasks have made extensive use of it. Salient object detection seeks to locate and extract from an image the objects that are most visually striking and significant.

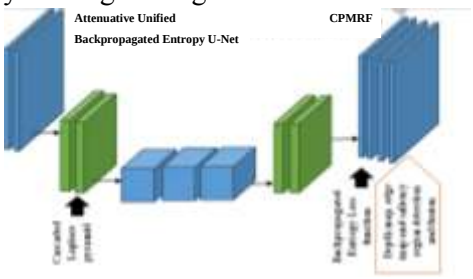


Fig. 2. U-Net's Design of the proposed model

Fig. 2 shows the design of the U-Net for the proposed Attenuative Unified U-Net with Conditional PatchMatch Algorithm for Edge Preserving Salient Object Detection model. The input is send to the U-Net input layer first, then the attention mechanism is operated to apply weights. Next the Cascaded Laplace Pyramid is incorporated to handle depth, edge, and salient regions. Next, the network is motivated to create the Backpropagated Entropy Loss Function to generate an overall fused image output. In the U-Net's output layer, for the depth discontinuity problem, the CPMRF approach combines the benefits of contextual modeling with the efficiency of Patch Match.

The Attention Mechanism, which learns to apply weights to various spatial locations or channels of the feature maps, is merged with the Attenuative Unified Backpropagated Entropy U-Net approach to emphasize regions or features that are regarded most relevant for edge and salient region detection. The attention mechanism is useful in encoder-decoder architectures such as the U-net since it gives localized classification information as opposed to global classification. This enables various network segments in U-net to concentrate on segmenting distinct objects. Additionally, the network is trained to recognize specific items in a picture with correctly labeled training data. The attention gate operates by applying a function that weights the feature map based on each class, tuning the network to concentrate on a specific class and, as a result, focus on specific objects in an image. The attention mechanism is described by equation (3),

$$q_{att}^{l} = \psi^{T}\left(\sigma_1\left(W_y^T y_i^l + W_g^T g_i + b_g\right)\right) + b_\psi \quad (3)$$

where y^l is the features from the contracting path and g is the gating signal.

Then the Cascaded Laplace pyramid is incorporated into the Unified U-Net's design to handle depth, edge, and salient regions. This allows the network to process the image at multiple scales. Then, by adding extra output channels to the decoder network, which effectively predict the depth and edge maps with the salient region at different scales, the depth and edge maps with the salient region are created simultaneously from a single input image based on the contextual information and depth cues. The Laplace Operator is defined in equation (4),

$$l_i^n = -\sum_{i=0}^{size(X)}(\Delta X_i \log(\Delta P_i) + (1 - \Delta X_i) log(1 - \Delta P_i)) \quad (4)$$

where, Δ signifies the Laplace operation. The Laplace Operator is initially utilized to determine the boundaries of the saliency map and ground truth using the operation to obtain more accurate salient object boundaries.

Encoder:

Input: X (Original Input Image)
Encoder Layer 1: H_1=ConvBlock (X,64)
Pooling Layer 1: P_1=MaxPooling (H_1)
Laplace Pyramid Level 1: L_1=LaplacePyramid (X,32)
Concatenate 1: C_1=Concatenate (P_1, L_1)
Encoder Layer 2: H_2=ConvBlock (C_1,128)
Pooling Layer 2: P_2=MaxPooling (H_2)
Laplace Pyramid Level 2: L_2=LaplacePyramid ($P1$,16)
Concatenate 2: C_2=Concatenate (P_2, L_2)
… (Continue the encoder structure)

Decoder:

… (Continue the decoder structure)
UpSampling Layer 3: U_3=UpSampling(C_2)
Decoder Layer 3: D_3=ConvBlock(U_3,64)
Output Layer: Y=Conv2D (D_3, num_classes, (1,1), activation='*softmax*')

Here, ConvBlock represents a series of convolutional layers, MaxPooling represents the max-pooling operation, LaplacePyramid represents the Laplace pyramid processing, and Concatenate represents the concatenation operation.

The network is then encouraged to produce an overall fused image output that is both accurate and confident by developing the Backpropagated Entropy Loss Function. The overall entropy loss is processed as the weighted sum of the individual backpropagated entropy losses. To diminish the backpropagated entropy loss, the model's weights are updated via backpropagation during training. Equation (5) shows the backpropagation process,

$$W_i = W_{i-1} - \alpha(\frac{\partial_{error}}{\partial_{W_i}}) \quad (5)$$

Where, W_i is the new weight, W_{i-1} is the out weight, and α is the learning rate.

In the U-Net's final output layer, the depth map, edge map, and salient area map are combined to produce an overall fused image. Let Y be the U-Net's output, Y_{depth} be the depth map, Y_{edge} be the edge map, and $Y_{salient}$ be the salient area map, which is defined in equation (6)(7)(8).

$$Entropy\ Loss_{depth} = -\frac{1^{N\,C}}{N_{i=1\ j=1}} Y_{depth}(i,j).\log(Y(i,j)) \quad (6)$$

$$Entropy\ Loss_{edge} = -\frac{1^{N\,C}}{N_{i=1\ j=1}} Y_{edge}(i,j).\log(Y(i,j)) \quad (7)$$

$$Entropy\ Loss_{salient} = -\frac{1^{N\,C}}{N_{i=1\ j=1}} Y_{salient}(i,j).\log(Y(i,j)) \quad (8)$$

where N is the count of pixels, C is the count of classes, and $Y(i,j)$ is the predicted probability for class j at pixel i. The overall entropy loss is defined in equation (9),

$$Overall\ Entropy\ Loss = -\frac{1^{N\,C}}{N_{i=1\ j=1}} Y_{d+e+s}(i,j).\log(Y(i,j).\ w_{d+e+s} \quad (9)$$

where, w_{d+e+s} is the weights assigned to the depth, edge, and salient maps, respectively and Y_{d+e+s} is the probability of depth, edge, and salient maps.

The backpropagation algorithm is then used to update the model's weights to minimize the overall entropy loss. This involves computing the gradients of the overall entropy loss regarding the model's parameters and adjusting the parameters in the direction that reduces the loss.

$$Y = w_{depth}.Y_{depth} + w_{edge}.Y_{edge} + w_{salient}.Y_{salient}$$
$$Y = w_{d+e+s}.Y_{d+e+s} \quad (10)$$

Equation (10) represents a linear combination of the depth, edge, and salient maps with their respective weights to produce the final fused image Y. Note that the weights w_{depth}, w_{edge}, and $w_{salient}$ determine the influence of each map in the final output.

For the depth discontinuity problem, the CPMRF approach is described; this approach combines the benefits of contextual modeling with the efficiency of Patch Match. The PatchMatch algorithm iteratively propagates matches from nearby patches. The matching cost C for a pair of patches (p, q) is computed, and the algorithm updates the matches based on the cost. The matching cost could involve measures like pixel differences or feature similarities. The matching cost is used to iteratively update the best-matching patches. After obtaining

initial PatchMatch results, CRFs are applied to refine the results by considering contextual information. The CRF model aims to minimize this energy function by adjusting the depth values. Equation (11) explains the CPMRF process,

$$E(X,p,q) = {}_{i}\Psi_{data}(X_i) + {}_{i,}j\Psi_{smooth}(X_i,X_j) + C(p,q) \quad (11)$$

where, X is the set of random variables corresponding to the depth values of pixels, $\Psi_{data}(X_i)$ is the data term that encourages the depth values to be similar to the PatchMatch results, and $\Psi_{smooth}(X_i,X_j)$ is the smoothness term that encourages smooth transitions between neighboring pixels.

The CPMRF algorithm integrates the PatchMatch results and CRF refinement. The refined depth map D_r is represented in equation (12) as,

$$D_r = {}_{i}\Psi_{data}(X_i) + {}_{i,}j\Psi_{smooth}(X_i,X_j) + C(p,q).D_i \quad (12)$$

where D_i is the initial depth map obtained from the PatchMatch algorithm. This objective combines the matching cost from the PatchMatch results and the smoothness term from the CRF model. The refined depth map D_r obtained from the CPMRF algorithm is used to enhance the U-Net's output. This involve combining the refined depth map with the U-Net's predictions to enhance the accuracy of salient object prediction. The final output Y_{output} is shown in equation (13),

$$Y_{output} = \left({}_{i}\Psi_{data}(X_i) + {}_{i,}j\Psi_{smooth}(X_i,X_j) + C(p,q).D_i\right) + (w_{d+e+s}.Y_{d+e+s})$$
$$Y_{output} = (D_r + Y) \quad (13)$$

Fig.3. Flowchart of the proposed method

Fig .3 describes the flowchart of the proposed method. First, the Attenuative Unified Backpropagated Entropy U-Net is introduced which integrates an Attention mechanism to assign weights to different spatial locations. Then, the Cascaded Laplace pyramid is incorporated with the U-Net which captures features at different sizes. Then, the Backpropagated Entropy Loss Function is defined to encourage the network to produce an overall fused image output that is both accurate and confident. Next, the CPMRF is introduced which combines the benefits of Patch Match efficiency and CRFs' contextual modeling.

4. Result and Discussion

This section contains a comprehensive analysis of the implementation outcomes, evaluates the performance of the proposed system, and includes a comparison segment to demonstrate the suitability of the proposed system for Attenuative Unified U-Net with Conditional PatchMatch Algorithm in Edge Preserving Salient Object Detection.

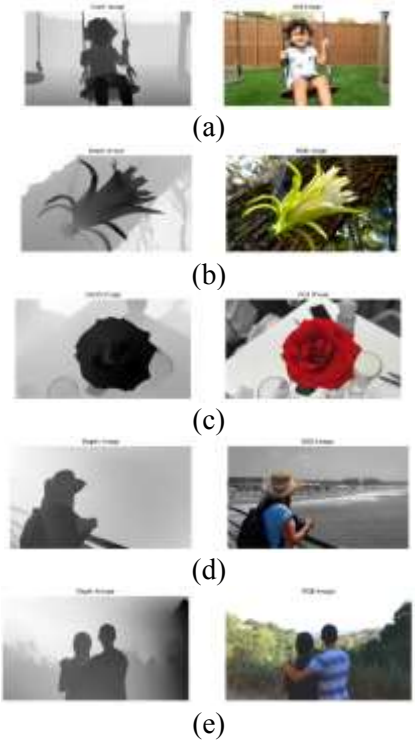
4.1 System configuration

The proposed system is experimented with in MATLAB, and this section gives a full description of the implementation findings and the proposed system's performance, as well as a comparative analysis to confirm that the proposed system works well.

Software	: MATLAB
OS	: Windows 10
Processor	: Intel i5
RAM	: 8GB

4.2 Simulated outcomes of the proposed model

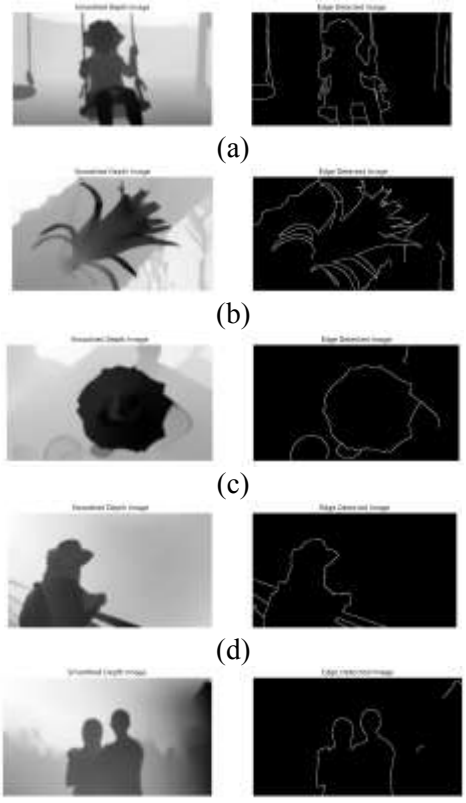
This section highlights the simulated output of the proposed model, Attenuative Unified U-Net with the Conditional PatchMatch Algorithm for Edge Preserving Salient Object Detection, starting from the initial setup.



(a) (b) (c) (d) (e)

Fig.4. Inputs taken from the dataset.

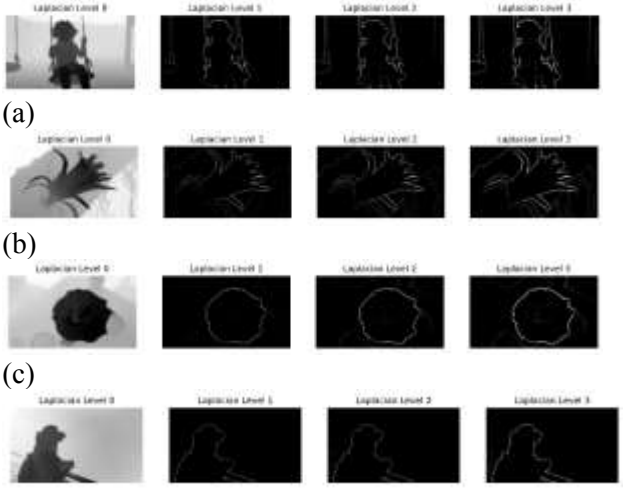
Fig. 4 shows the inputs taken from the datasets. Here the depth and the RGB image are taken together from the RGB-D dataset. Totally five different input samples are taken for the implementation to get better results.



(a) (b) (c) (d) (e)

Fig.5. Edge Detected image

Fig.5 displays the edge detected image of the given inputs. The edge detected image is obtained from the smoothed depth image. This process is done by the Attenuative Unified Backpropagated Entropy U-Net method, which integrates an Attention Mechanism that learns to allot weights to diverse spatial locations or channels of the feature maps, highlighting regions or features that are deemed most relevant for edge and salient region detection.

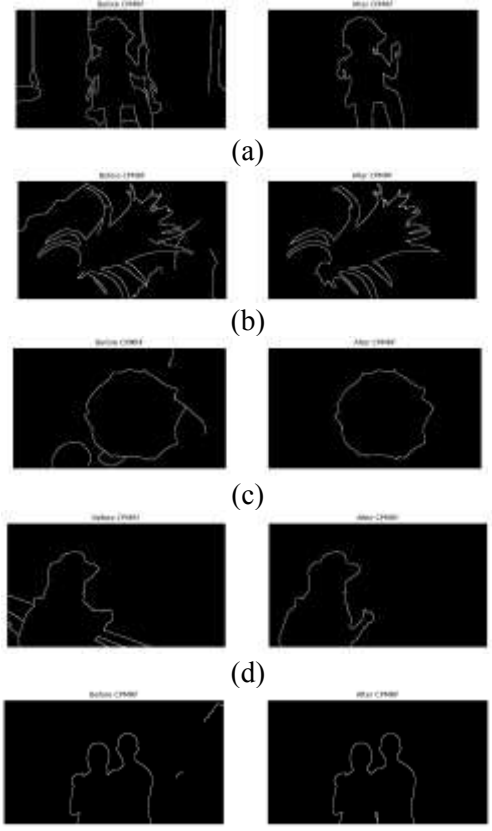


(a) (b) (c) (d)

(e)

Fig.6. Laplacian Levels

Fig.6 demonstrates the Laplacian levels of the given samples. A variety of information types can be handled by the Unified U-Net, which includes the Cascaded Laplace pyramid, which breaks down an image into a number of band-pass filtered versions by combining the series of Laplacian pyramids. Totally, four laplacian levels were taken place from level 0 to level 3, which helps to obtain clear boundary images.



(a) (b) (c) (d) (e)

Fig.7. CPMRF Algorithm

Fig.7 illustrates the Conditional PatchMatch Random Field Algorithm (CPMRF) applied to the samples taken. The CPMR provides a solution to the depth discontinuity problem. PatchMatch repeatedly propagates matches from surrounding patches to discover comparable patches in the remaining section of the image. By considering the connections between the matched patches and their surrounding areas, CRF enhances the PatchMatch results.

(a)

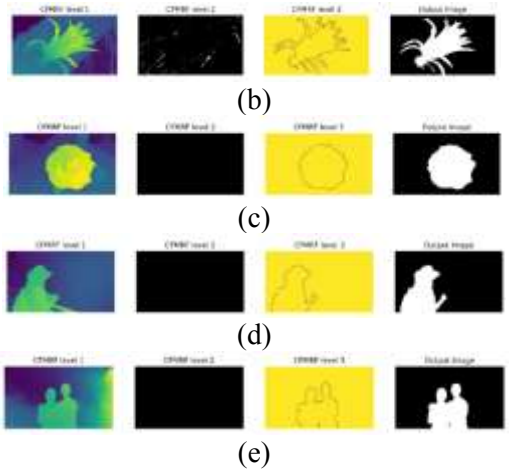


Fig. 8. Output

Fig. 8 elucidates the outcome of the sample images after the completion of each step. Three CPMRF levels were taken from level 1 to level 3 and finally, the output is displayed. The proposed method's output is better and more effective when compared with other existing methods.

4.3 Performance metrics of the proposed system

In this section, a detailed clarification of the efficacy of the suggested technique and the achieved outcome were explained.

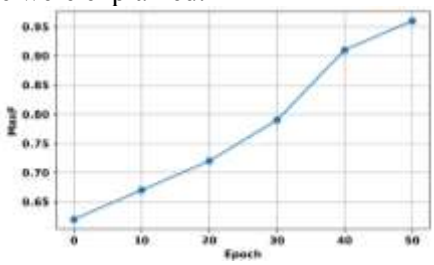


Fig. 9. MaxF of the proposed system

The MaxF of the proposed system for a diverse count of epochs has been shown in Fig. 9. The MaxF of the proposed system gains a maximum value of 0.96 when the count of epochs is increased to 50 and gains a minimum value of 0.62 when the count of epochs is abridged. In salient object recognition, the combination of an Attenuative Unified U-Net with a Conditional PatchMatch Algorithm probably improves MaxF by strengthening the network's capacity to extract pertinent features, adjust to changing circumstances, and maintain edges efficiently.

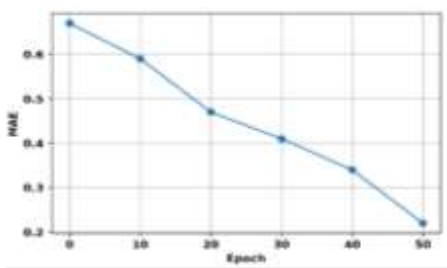


Fig.10. MAE of the proposed system

The MAE of the proposed system achieves a minimum value of 0.21 when the number of epochs is increased to 50 and attains a maximum value of 0.64 when the number of epochs is reduced. This is shown in Figure 10, where the MAE of the proposed system varies with the number of epochs. When it comes to salient object recognition, attenuative backpropagated entropy U-Net techniques help reduce MAE. Accurate output images can be obtained using this. The MAE error rate in salient object detection is decreased by attenuating non-salient regions.

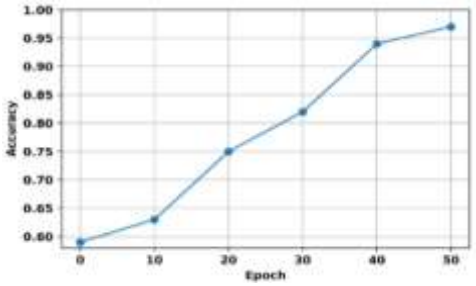


Fig. 11. Accuracy of the proposed system

Fig. 11 displays how accurate the suggested method is at changing the number of epochs. When more epochs are used, the accuracy of the suggested system reaches a maximum value of 0.9625 and a value of 0.83 when 30 epochs are used. The effectiveness of CPMRF in locating estimated nearest neighbors is well established. Incorporating this approach allows information to propagate throughout the image quickly and effectively, improving accuracy.

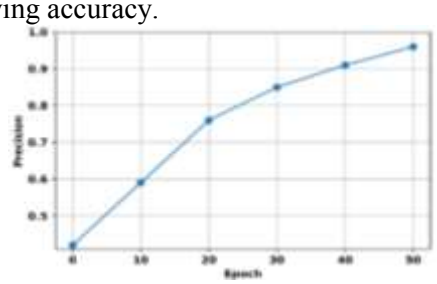


Fig. 12. Precision of the proposed system

Fig. 12 illustrates the performances of the precision of the proposed model. When the epoch value is 1 it achieves the maximum precision value of 0.96 and while the iteration value is 50, it achieves the precision value of 0.83 when the epoch is 30. The model attains an equilibrium between capturing minute features and taking into account the wider environment by employing the CPMRF algorithm for contextual refinement and the U-Net architecture for edge preservation, which both contribute to increased precision.

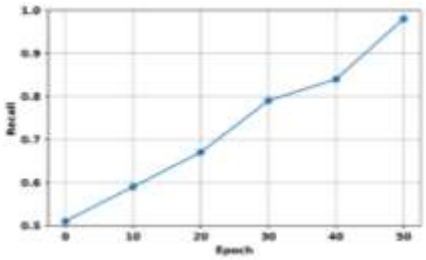


Fig. 13. Recall of the proposed system

The Fig. 13 illustrates the recall of the proposed system when the number of epochs is varied. The proposed method achieves a maximum recall of 0.9610 when the number of epoch is 50 and attains a recall of 0.75 when the number of epochs is 30. Attenuative Backpropagated Entropy U-Net mechanisms aid in lessening the effects of particular elements or qualities. This is useful when the model has to concentrate on particular areas or details inside the image. Attenuating non-salient regions improves recall in salient object detection by reducing false negatives.

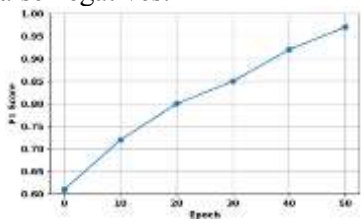


Fig. 14. F1 Score of the proposed system

The Fig. 14 depicts the F1 Score of the proposed system for altering the number of epoch. When the number of epochs is 50, the F1 score of the proposed system reaches a maximum of 0.9735 and achieves a F1 score of 0.85 when the number of epochs is 30. The Attenuative Unified U-Net and Conditional PatchMatch pay attention to prominent regions and preserve edges, and also refines the segmentation findings, which helps to improve the F1 Score.

4.4 Comparison of Proposed Model with Previous Models

This section illustrates the outcomes based on several metrics, emphasizing the effectiveness of the proposed model by comparing it with existing methodologies. The comparisons are made from previous techniques with various metrics: accuracy, precision, recall, F1 score, MaxF, and mean absolute error (MAE). Comparisons are made with the existing techniques such as Pixel-wise contextual attention Network (PiCANet), Edge Guidance Network (EGNet), Cascaded Partial Decoder (CPD), and Deeply Supervised Structure (DSS).

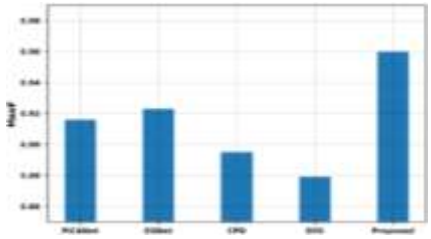


Fig. 15. Comparison of the MaxF of the proposed model

Fig. 15 exemplifies the comparison of the MaxF of the proposed model with existing models. The existing models such as PiCANet, EGNet, CPD, and DSS gains a MaxF value of 0.915, 0.921, 0.894, and 0.88 respectively. Computed with existing models the proposed model gains a MaxF of 0.96.

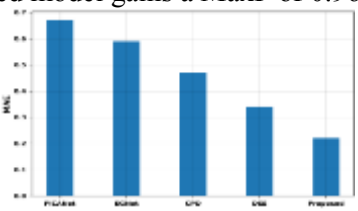


Fig. 16. Comparison of the MAE of the proposed model

The comparison of the suggested model's MAE with that of current models is shown in Figure 16. The MAE values of the current models, including PiCANet, EGNet, CPD, and DSS, are 0.65, 0.54, 0.35, and 0.65, respectively. When compared to current models, the suggested model has a mean average error of 0.21.

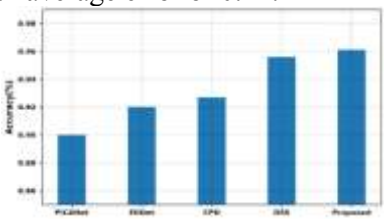


Fig. 17. Comparison of the Accuracy of the proposed model

Figure 17 illustrates the comparison of the accuracy of the proposed model with existing models. The existing models such as PiCANet, EGNet, CPD, and DSS achieves an accuracy value of 0.90, 0.92, 0.925, and 0.953 respectively. Compared with existing models the proposed model achieves an accuracy of 0.9625.

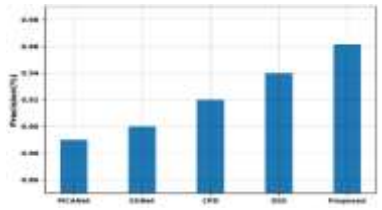


Fig. 18. Comparison of the Precision of the proposed model

Figure 18 illustrates the comparison of the precision of the proposed model with existing models. The existing models such as PiCANet, EGNet, CPD, and DSS achieves a precision value of 0.89, 0.90, 0.92, and 0.94 respectively. Compared with existing models the proposed model achieves a precision of 0.96.

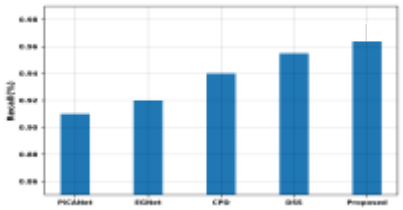


Fig. 19. Comparison of the Recall of the proposed model

Fig. 19 illustrates the comparison of the recall of the proposed model with existing models. The existing models such as PiCANet, EGNet, CPD, and DSS achieves a recall value of 0.91, 0.92, 0.94, and 0.952 respectively. Compared with existing models the proposed model achieves a recall of 0.9610.

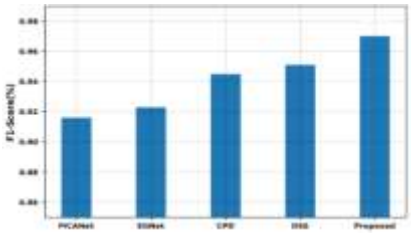


Fig. 20. Comparison of the F1 Score of the proposed model

Fig. 20 illustrates the comparison of the F1 score of the proposed model with existing models. The existing models such as PiCANet, EGNet, CPD, and DSS achieves a F1 score value of 0.913, 0.921, 0.945, and 0.95 respectively. Compared with existing models the proposed model achieves a F1 score of 0.9735.

Table 1: Performance comparison of different methods

Models	MaxF	MAE	Accuracy	Precision	Recall	F1 Score
PiCANet	0.915	0.65	0.90	0.89	0.91	0.913
EGNet	0.921	0.54	0.92	0.90	0.92	0.921
CPD	0.894	0.44	0.925	0.92	0.94	0.945
DSS	0.88	0.35	0.953	0.94	0.952	0.95
Proposed	**0.96**	**0.21**	**0.9625**	**0.96**	**0.9610**	**0.9735**

Table 1 shows the performance comparison table of different models. The proposed method yielded the highest MaxF of 0.96, while the existing methods such as PiCANet, EGNet, CPD, and DSS attained the MaxF of 0.915, 0.921, 0.894, and 0.88 respectively. The proposed method yielded the lowest MAE of 0.21, while the existing methods such as PiCANet, EGNet, CPD, and DSS attained the MAE of 0.65, 0.54, 0.44, and 0.35 respectively. The proposed method yielded the highest accuracy of 0.9625, while the existing methods such as PiCANet, EGNet, CPD, and DSS attained the accuracy of 0.90, 0.92, 0.925, and 0.953 respectively. The proposed method yielded the highest precision of 0.96, while the existing methods such as PiCANet, EGNet, CPD, and DSS attained the precision of 0.89, 0.90, 0.92, and 0.94 respectively. The proposed method yielded the highest recall of 0.9610, while the existing methods such as PiCANet, EGNet, CPD, and DSS attained the recall of 0.91, 0.92, 0.94, and 0.952 respectively. The proposed method yielded the highest F1 Score of 0.9735, while the existing methods such as PiCANet, EGNet, CPD, and DSS attained the F1 Score of 0.913, 0.921, 0.945, and 0.95 respectively. The comparison results confirm the superiority of the presented model over the existing methods.

Overall, this section gives the detailed explanation of the results obtained by the proposed method which includes the system configuration, simulation output, performance of the proposed method, and the comparison of the proposed method with the existing methods. As a result, the generated model proved to be the most exact and precise compared to the other existing methods.

5. Conclusion

A simple yet effective Attenuative Unified U-Net with Conditional PatchMatch Algorithm for Edge Preserving Salient Object Detection was proposed here to overcome the issues in edge-preserving. First, Attenuative Unified Backpropagated Entropy U-Net, incorporated an Attention Mechanism by which the feature maps' various spatial locations or channels were trained to receive weights. Then, the Unified U-Net, which combined the Cascaded Laplace pyramid, combined the sequence of Laplacian pyramids to deconstruct an image into a series of band-pass filtered copies, was designed to handle many forms of information. The network was then encouraged to generate an overall fused image output that was both accurate and confident by defining the Backpropagated Entropy Loss Function. Finally, for the depth discontinuity

problem, the CPMRF algorithm, which combined the advantages of Patch Match efficiency and contextual modeling provided by CRFs. PatchMatch found similar patches in the remaining portion of the image by repeatedly propagating matches from nearby patches. CRF, which improve the PatchMatch results by taking into account the connections between the matched patches and the areas around them. Thus, the produced model turned out to be the most accurate and precise of all the previously developed models, scoring 0.9625 of accuracy, 0.96 of precision and yielded recall, MaxF, MAE and F1 score of 0.9610, 0.96, 0.21 and 0.9735 respectively.

References

[1] M. Zhuge, D. P. Fan, N. Liu, D. Zhang, D. Xu, and L. Shao, "Salient object detection via integrity learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*. vol. 45 no. 3, pp. 3738-3752, 2022.

[2] Y. Zhang, L. Zhang, W. Hamidouche, and O. Deforges, "A fixation-based 360 benchmark dataset for salient object detection," *In 2020 IEEE International Conference on Image Processing (ICIP*), pp. 3458-3462 IEEE. 2020 October.

[3] Z. Chen, H. Zhou, J. Lai, L.Yang, and X. Xie, "Contour-aware loss: Boundary-aware learning for salient object segmentation," *IEEE Transactions on Image Processing,* vol. 30, pp. 431-443, 2020.

[4] X. Huang and Y.J. Zhang. "Fast video saliency detection via maximally stable region motion and object repeatability," *IEEE Transactions on Multimedia,* vol. 24, pp. 4458-4470, 2021.

[5] F. An, J. Wang, and R. Liu, "Pedestrian Re-Identification Algorithm Based on Attention Pooling Saliency Region Detection and Matching," *IEEE Transactions on Computational Social Systems*, 2023.

[6] Z. Tu, Z. Li, C. Li, Y. Lang, and J. Tang, "Multi-interactive dual-decoder for RGB-thermal salient object detection," *IEEE Transactions on Image Processing,* vol. 30, pp. 5678-5691, 2021.

[7] J. Zhang, Y. Liu, S. Zhang, R. Poppe, and M. Wang, "Light field saliency detection with deep convolutional networks," *IEEE Transactions on Image Processing,* vol. 2, pp. 4421-4434, 2020.

[8] L. Xu, and X. Xu, "RGB-D Visual Saliency Detection Algorithm based on Information Guided and Multimodal Feature Fusion," *IEEE Access,* 2023

[9] M.M. Cheng, S.H. Gao, A. Borji, Y.Q. Tan, Z. Lin, and M. Wang, "A highly efficient model to study the semantics of salient object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 44, no. 11, pp. 8006-8021, 2021.

[10] R. Cong, N. Yang, C. Li, H. Fu, Y. Zhao, Q. Huang, and S.K wong, "Global-and-local collaborative learning for co-salient object detection," *IEEE transactions on cybernetics,* vol. 53, no. 3, pp. 1920-1931, 2022.

[11] P. Chen, and B. Li, "RGB-D Salient Object Detection via Joint Learning and Multi-feature Fusion", *In 2022 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML), pp.* 547-552. IEEE, 2022 October.

[12] X. Zhou, H. Fang, Z. Liu, B. Zheng, Y. Sun, J. Zhang, and C. Yan, "Dense attention-guided cascaded network for salient object detection of strip steel surface defects," *IEEE Transactions on Instrumentation and Measurement,* vol. 71, pp. 1-14, 2021.

[13] X. Zhu, J. Wu, and L. Zhu, "RGB-D Saliency Detection based on Cross-Modal and Multi-Scale Feature Fusion," *In 2022 34th Chinese Control and Decision Conference (CCDC),* pp. 6154-6160 IEEE, 2022 August.

[14] G. Li, Z. Liu, M. Chen, Z. Bai, W. Lin, and H. Ling, "Hierarchical alternate interaction network for RGB-D salient object detection," *IEEE Transactions on Image Processing,* vol. 30, pp. 3528-3542, 2021.

[15] Q. Chen, Z. Zhang, Y. Lu, K. Fu, and Q. Zhao, "3-d convolutional neural networks for rgb-d salient object detection and beyond," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[16] S. Zhou, J. Wang, L. Wang, J. Zhang, F. Wang, D. Huang, and N. Zheng, "Hierarchical and interactive refinement network for edge-preserving salient object detection," *IEEE Transactions on Image Processing,* vol. 30, pp. 1-14, 2020.

[17] Z. Tu, Y. Ma, C. Li, J. Tang, and B. Luo, "Edge-guided non-local fully convolutional network for salient object detection," *IEEE transactions on circuits and systems for video technology,* vol. 31, no. 2, pp. 582-593, 2020.

[18] H. Wen, C. Yan, X. Zhou, R. Cong, Y. Sun, B. Zheng, J. Zhang, Y. Bao, and G. Ding, "Dynamic selective network for RGB-D salient object detection," *IEEE Transactions on Image Processing,* vol. 30, pp. 9179-9192, 2021.

[19] L. Sun, Z. Chen, Q.J. Wu, H. Zhao, W. He, and X. Yan, "AMP Net: Average-and max-pool networks for salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 11, pp. 4321-4333, 2021.

[20] X. Li, D. Song, and Y. Dong, "Hierarchical feature fusion network for salient object detection," *IEEE Transactions on Image Processing,* vol. 29, pp. 9165-9175, 2020.

[21] Z. Wu, S. Li, C. Chen., H. Qin, and A. Hao, "Salient object detection via dynamic scale routing," *IEEE Transactions on Image Processin*g, vol. 31, pp. 6649-6663, 2022.

[22] D. Kang, S. Park, and J. Paik, "Coarse to fine: Progressive and multi-task learning for salient object detection," *In 2020 25th International Conference on Pattern Recognition (ICPR),* pp. 1491-1498 IEEE, 2021 January.

[23] B.V. Lad, M.F. Hashmi, and Keskar A.G, "Boundary preserved salient object detection using guided filter based hybridization approach of transformation and spatial domain analysis," *IEEE Access* vol. 10, pp. 67230-67246, 2022.

[24] M. Xu, P. Fu, B. Liu, and J. Li. "Multi-stream attention-aware graph convolution network for video salient object detection. *IEEE Transactions on Image Processing*, vol. 30, pp. 4183-4197, 2021.

[25] Y. Liu, X.Y. Zhang, J.W. Bian, L. Zhang, and M.M Cheng, "SAM Net: Stereoscopically attentive multi-scale network for lightweight salient object detection," *IEEE Transactions on Image Processing,* vol. 30, pp. 3804-3814, 2021.

[26] L. Zhang, J. Wu, T. Wang, A. Borji, G. Wei, and H. Lu, "A multistage refinement network for salient object detection," *IEEE Transactions on Image Processing,* vol. 29: pp. 3534-3545, 2020.

[27] W. Zhou, J. Wu, J. Lei, J. N. Hwang, and L. Yu, "Salient object detection in stereoscopic 3D images using a deep convolutional residual autoencoder," *IEEE Transactions on Multimedia*, vol. 23, pp. 3388-3399, 2020.

[28] X. Zhou, H. Fang, X. Fei, R. Shi, and J. Zhang, "Edge-aware multi-level interactive network for salient object detection of strip steel surface defects," *IEEE Access*, vol. 9, pp. 149465-149476, 2021.

[29] G. Li, Z. Liu, X. Zhang, and W. Lin, "Lightweight salient object detection in optical remote-sensing images via semantic matching and edge alignment," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 61, pp. 1-11, 2023.

[30] L. Wang, R. Chen, L. Zhu, H. Xie, and X. Li, "Deep sub-region network for salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 2, pp. 728-741, 2020.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.