SalFAU-Net: Saliency Fusion Attention U-Net for Salient Object Detection

KASSAW ABRAHAM MULAT
ZHENGYONG FENG
TEGEGNE SOLOMON ESHETIE
AHMED ENDRIS HASEN

Salient object detection (SOD) remains an important task in computer vision, with applications ranging from image segmentation to autonomous driving. Fully convolutional network-based methods have made remarkable progress in visual saliency detection over the last few decades. However, these methods have limitations in accurately detecting salient objects, particularly in challenging scenes with multiple objects, small objects, or objects with low resolutions. To address this issue, we proposed a Saliency Fusion Attention U-Net (SalFAU-Net) model, which incorporates an saliency fusion module into each decoder block of the attention U-net model to generate saliency probability maps from each decoder block. SalFAU-Net employs an attention mechanism to selectively focus on the most informative regions of an image and suppress nonsalient regions. We train SalFAU-Net on the DUTS dataset using a binary cross-entropy loss function. We conducted experiments on six popular SOD evaluation datasets to evaluate the effectiveness of the proposed method. The experimental results demonstrate that our method, SalFAU-Net, achieves competitive performance compared to other methods in terms of mean absolute error, *F*-measure, *S*-measure, and *E*-measure.

K. A. Mulat and T. S. Eshetie are with the School of Computer Science, China West Normal University, Nanchong, 637002 Sichuan, China (e-mail: abraham88@stu.cwnu.edu.cn; solomoneshetie8@gmail.com).

Z. Feng is with the School of Electronic and Information Engineering, China West Normal University, Nanchong, 637002 Sichuan, China (e-mail: zhyfeng@cwnu.edu.cn).

A. E. Hasen is with the Paul C. Lauterbur Research Center, Shenzhen Institute of Advanced Technology, Chinese Academy of Science, Shenzhen, Guangdong 518055, China (e-mail: ahmeduestc53@outlook.com).

(Corresponding author. Kassaw Abraham Mulat.)

I. INTRODUCTION

Salient object detection (SOD), also referred to as visual saliency detection, is detecting the most noticeable, unique, and visually distinct objects or regions in a scene that attract the human eye [3]. The human visual perception system has an exceptional ability to rapidly recognize and focus its attention toward visually unique and prominent objects or regions within scenes [37]. This innate capability has captivated the interest of many researchers in the field of computer vision, where the aim is to simulate this process based on the psychological and biological properties of the human visual attention system. The goal is to identify prominent objects in images and videos that hold significant importance and valuable information.

Given the diverse applications of SOD in various domains of computer vision, it plays a crucial role as a preprocessing step in tasks like image segmentation [1], [7], [12], [31], object detection [4], [35], [33], image captioning [43], autonomous driving [29], and augmented reality [8]. Numerous visual saliency detection methods have been proposed. These methods aim to distinguish the most unique foreground images from less significant backgrounds. While traditional saliency detection approaches rely on low-level heuristic visual features, these methods often fail to detect salient objects in challenging scenes. Recently, deep learning methods, particularly convolutional neural networks (CNNs), have exhibited exceptional efficacy across diverse computer vision tasks, including saliency detection. In contrast to traditional methods, CNN-based methods have made remarkable advancements by harnessing advanced semantic features [16].

Due to the significant impact of representative features on algorithms performance, it is beneficial to investigate models that leverage multilevel features and contextual information to enhance saliency detection. Furthermore, despite the introduction of end-to-end models based on fully convolutional networks (FCNs), there remains significance in incorporating and advancing conventional FCN models like U-Net [34] and its variate for the task of saliency detection. One of the variants of U-Net that is well known for its efficacy in medical image segmentation is the Attention U-Net network [28], which selectively focuses on relevant regions of the input image by integrating attention mechanisms into its architecture, which improves the model's ability to capture intricate patterns and important features. The attention mechanism facilitates improved performance in tasks such as image segmentation. Drawing on its success in medical image segmentation, this study explores the application of Attention U-Net for saliency detection tasks. We added a saliency fusion module (SFM) to each decoder block of the network. This module allows us to generate saliency maps effectively, which we then concatenate with each decoder's side output saliency map to get the final saliency map. The

attention gate (AG) module in the proposed method helps the model learn to focus on salient features with varying sizes and shapes. In this way, SalFAU-Net has the ability to suppress irrelevant regions from an input image while emphasizing the features that are most important for saliency detection. To summarize, the main contributions of this paper are as follows:

(1) We proposed a Saliency Fusion Attention U-Net (SalFAU-Net) for the task of visual saliency detection.

(2) SFM is added to each decoder block of the network to generate a saliency map from each decoder, and these saliency maps are concatenated together to obtain the ultimate visual representation that highlights the most important areas or objects in an image.

(3) We conducted experiments on six publicly available challenging SOD datasets, and the results demonstrate the effectiveness of SalFAU-Net for the task of visual saliency detection.

II. RELATED WORKS

Generally, saliency detection methods can be classified into two categories. These are traditional methods and deep learning-based methods. Traditional methods are based on low-level heuristic visual features such as contrast, location, and texture. Most of these methods are unsupervised or semi-supervised. Examples of traditional saliency detection methods include those based on local contrast [22], global contrast [44], background-ness prior [45], center prior [42], objectness prior [20], and others. These methods achieved good results in uncomplicated images or scenarios featuring solitary objects. Nonetheless, these methods failed to detect salient objects that are in complex scenes, low resolutions, or scenes with multiple salient objects. This limitation arises from their reliance on low-level features, which prove inadequate for addressing the complexities introduced by such challenging visual contexts.

Recently, deep learning-based methods, particularly CNNs, have demonstrated remarkable performance across diverse computer vision tasks, including image classification [17], semantic image segmentation [24], and object detection [41]. CNNs have the capability to learn rich and hierarchical representations of input data by extracting high-level semantic features. However, in SOD, both low-level and high-level features are important for developing good visual saliency detection models. The introduction of FCNs [25] has revolutionized the approach to end-to-end pixel-level saliency detection. Initially designed for semantic segmentation, FCN seamlessly combines the tasks of feature extraction and pixel label prediction in a single network structure composed of down-sampling and upsampling paths. Subsequently, numerous FCN-based visual saliency detection models have been proposed, including deep contrast learning (DCL) [19], aggregating multilevel convolutional feature frameworks (Amulet) [46], recurrent fully

convolutional networks (RFCN) [39], and deep uncertain convolutional features (UCF) [47]. These advancements have notably enhanced the effectiveness of algorithms designed for visual saliency detection. Nonetheless, exploring effective FCN-based models designed for different purposes is still beneficial. U-Net is one of the most widely used networks in medical image segmentation [34]. Following the success of U-Net, numerous network variations have been introduced for different tasks. One exemplary variant of U-Net is the Attention U-Net model, which is designed for pancreas image segmentation, has shown impressive results in other tissue and organ segmentation, benefiting from the AG module to focus on relevant and variable size regions in an image. Most FCN-based saliency models are based on plain U-Net and have achieved remarkable performance for saliency detection. In [32], Qin et al. proposed a two-level nested U-structure by using a residual U-block (RSU) as a backbone for visual saliency detection. Compared to many other networks that use pretrained networks as backbones, U-2-Net's RSU block increases architecture depth without significantly increasing computational costs while achieving competitive performance. In [14], Han et al. proposed a modified U-Net network for saliency detection, utilizing an edged convolution constraint. This variant effectively integrates features from multiple layers, reducing information loss and enabling pixel-wise saliency map prediction rather than patch-level prediction, which is common in CNN-based models.

Although these methods based on plain U-Net achieved remarkable performance for saliency detection, their performance can be boosted by incorporating different techniques into the encoder and decoder blocks of their architecture. Recently, attention mechanisms have shown remarkable results across various computer vision applications, encompassing saliency detection. In [21], Li et al. proposed a U-shape network with stacked layers incorporating channel-wise attention to extract the most important channel features and effectively utilize these features by integrating a parallel dilated convolution (PDC) module and a multilevel attention cascaded feedback (MACF) module.

In order to recurrently translate and aggregate the context features separately with various attenuation factors, Hu et al. [15] proposed a spatial attenuation context module. After that, the module carefully learned the weights to adaptively incorporate the collective contextual features. In [48], Zhang et al. proposed a novel approach to visual saliency detection that leverages attention mechanisms for refining saliency maps, incorporating bi-directional refinement for enhanced accuracy. The introduction of bi-directional refinement highlights the focus on comprehensive feature extraction and optimization. In [49], Zhao and Wu applied spatial attention (SA) and channel-wise attention (CA) to distinct aspects of the model. Specifically, SA was employed for low-level feature maps, while CA was incorporated into context-

aware pyramid feature maps. This strategic approach aims to direct the network's focus toward the most relevant features for the given sample. In [13], Gong *et al.* proposed an enhanced U-Net model incorporating pyramid feature attention, channel attention, and a pyramid feature extraction module to improve the performance of the U-Net backbone network.

In this research, we attempt to explore the applications of attention U-Net architecture in the realm of visual saliency detection. We added an SFM to each decoder of the network and concatenated their output to obtain the final saliency map. The AG module in the proposed method helps the model learn to focus on salient features with varying sizes and shapes. Thus, SalFAU-Net adeptly learns the capability to suppress irrelevant or undesirable regions within an input image while highlighting the most crucial and salient features essential for the task of saliency detection.

III. METHODOLOGY

In this section, we provide a detailed description of the architecture of our proposed method. This is followed by the network supervision, the datasets and evaluation metrics used, and the implementation details.

A. Architecture of SalFAU-Net

The proposed SalFAU-Net for visual saliency detection in this paper mainly consists of four parts: (1) a five-level encoder block, (2) a four-level decoder block, (3) an AG module, and (4) an SFM. Figure 1 shows the architecture of the proposed SalFAU-Net model. Compared with the Attention U-Net model proposed for pancreas image segmentation [28], we add an SFM to each decoder of the architecture and finally concatenate them together to obtain the final saliency map.

1) Encoder Block: Each encoder block consists of two convolutional layers, each followed by a batch normalization layer and ReLu activation function, which increases the number of feature maps from 3 to 1024. Max pooling with a stride of 2×2 is applied at the end of every block except the last block for downsampling, reducing the image size from 288×288 to 18×18. The encoder block progressively reduces the spatial resolution of feature maps while increasing the number of channels, capturing features at different scales.

2) Decoder Block: The decoder block is responsible for upsampling and generating salient maps. It consists of an up-sampling layer followed by two convolutional layers, batch normalization, and ReLU activation function. The decoder block is connected to the AG block through skip connections. Each decoder block reduces the number of feature maps by two while increasing the size of the spatial resolutions from 18×18 to 288×288. The goal is to recover spatial details lost during the downsampling in the encoder, facilitating precise localization and detection of salient objects.

3) AG Module: AGs have demonstrated remarkable effectiveness in capturing crucial regions, diminishing feature responses in irrelevant background areas, and eliminating the need to crop a region of interest (ROI) within an image. This is particularly important for the task of visual saliency detection. The integration of AGs into the conventional U-Net architecture enhances the model's ability to emphasize salient features transmitted through skip connections. Given a skip connection feature $F_s \in R^{C \times H \times W}$, where C is the number of channels, and H and W are the height and width of F, we first apply a convolution layer, batch norm, and ReLu activation function to obtain a key feature K, and let Q be the input from the previous layer or the gating signal obtained by applying a convolution layer followed by a batch
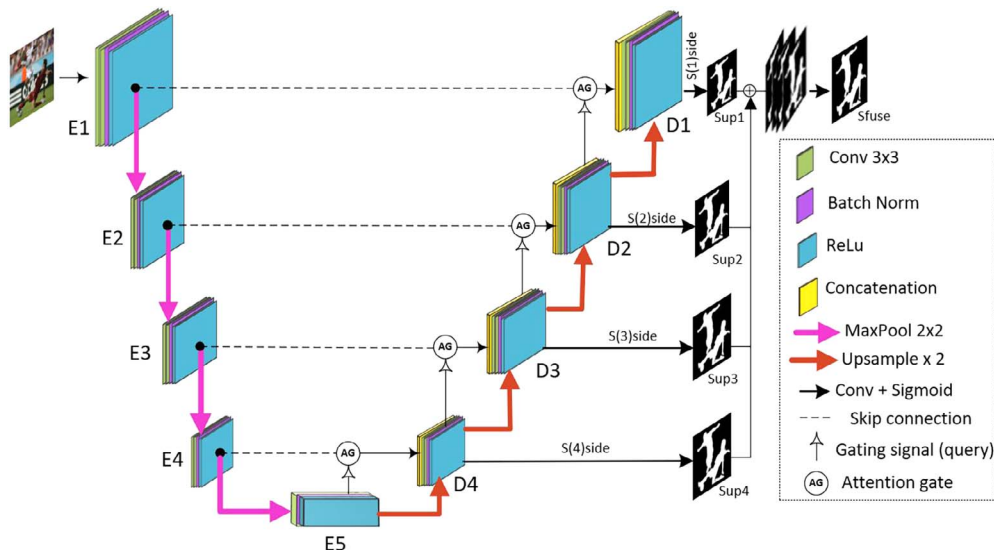

Figure 1.   Architecture of our proposed Saliency Fusion Attention U-Net (SalFAU-Net model).
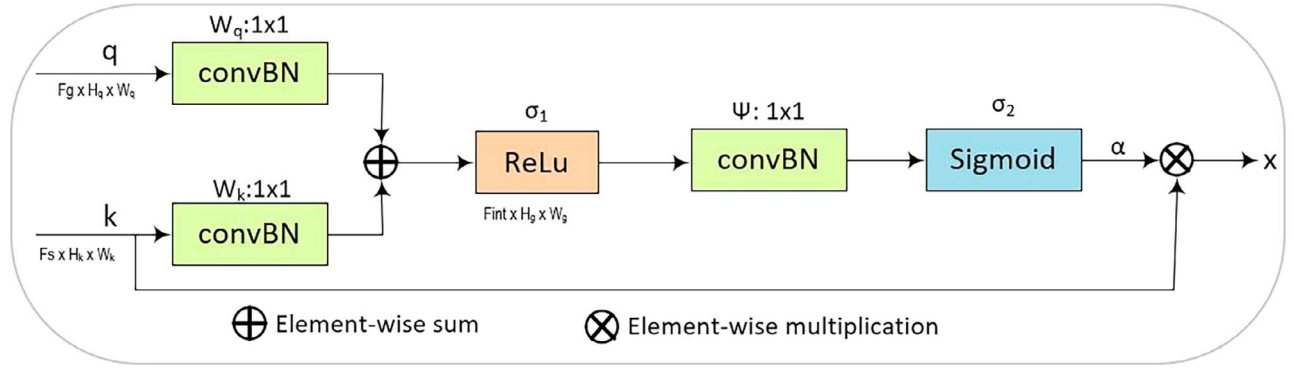
Figure 2. Diagram of the additive attention gate (AG) module.

normalization and relu activation to the input gate feature F_g. The attention coefficient α is obtained by applying a relu function to the element-wise sum of Q and K. The final attention coefficient value V is obtained by feeding a convolution layer, batch normalization, and sigmoid activation function to the attention coefficient α.

Finally, the attention coefficient value V and the skip connection feature map are multiplied element-wise to produce the final AG output $\hat{x}_{i,c}^l$, which is calculated as equation (3)

$$q_{att}^l = \Psi\left(\sigma_1\left(W_q^T \times q_i^l + W_k^T \times K_i + b_k\right)\right) + b_\Psi, \quad (1)$$

$$\alpha_i^l = \sigma_2\left(q_{att}^l\left(q_i^l, K_i; \Theta_{att}\right)\right), \quad (2)$$

$$\hat{x}_{i,c}^l = \alpha_i^l \cdot K_{i,c}, \quad (3)$$

where $\sigma_2(x_i, c) = \frac{1}{1+exp^{(-x_i,c)}}$ represents the sigmoid activation function. Thus, AG is defined by a set of parameter set θ_{att}, which includes linear transformations $W_k \in R^{F_l \times F_{int}}, W_q \in R^{F_g \times F_{int}}, \Psi \in R^{F_{int} \times 1}$ and bias terms $b_\Psi \in R, b_k \in R^{F_{int}}$. The linear transformation can be computed using channel-wise 1×1 convolutions for the input tensors.

4) Saliency Fusion Module: The saliency map fusion module serves as a pivotal component in generating saliency probability maps. Similar to the methodology in [32], our model undertakes a multistage approach. Initially, it generates four-side output saliency probability maps, denoted as $S(1)_{side}, S(2)_{side}, S(3)_{side}$, and $S(4)_{side}$, originating from the respective stages decoder1, decoder2, decoder3, and decoder4. This generation is facilitated by a 3×3 convolution layer, followed by a sigmoid activation function. Subsequently, the convolution outputs prior to sigmoid functions of these side-output saliency maps are upsampled to have the same size as the input image. The integration of these saliency maps is accomplished through a concatenation operation, followed by a 1×1 convolution layer and a sigmoid function. The result of this fusion process is the final saliency map S_{fuse} (depicted in the bottom right of Fig. 1).

Mathematically, the saliency probability maps at each stage are generated as follows:

$$S(i)_{side} = \sigma(\text{Conv}(i)(X)), \quad (4)$$

where i represents the stage (1, 2, 3, or 4), σ denotes the sigmoid function, $\text{Conv}(i)$ is the convolution operation at stage i, and X is the decoder's output feature map. The side outputs are then upsampled and concatenated to generate the final saliency map S_{fuse}:

$$S_{fuse} = \sigma(\text{Conv}_{fuse}(\text{Concat}(S(i)_{side}))), \quad (5)$$

where Concat represents the concatenation operation, Conv_{fuse} is the 1×1 convolution layer specific to the fusion process, and σ represents the sigmoid function.

B. Network Supervision

Loss functions play a significantly important role in optimizing a saliency detection model. One of the most widely employed loss functions for binary classification problems is the binary cross-entropy (BCE) loss [5]. For visual saliency detection, it measures the dissimilarity between the predicted saliency map and the ground truth in a binary classification setting.

We use a deep supervision approach similar to that in [32], which has demonstrated efficacy. Our training loss is formulated as follows:

$$\mathcal{L} = \sum_{m=1}^M w_{side}^m l_{side}^m + w_{fuse} l_{fuse} \quad (6)$$

The total loss comprises two components. The first component is the loss associated with the side-output saliency maps, denoted as l_{side}^m, where m represents the four supervision stages (Sup1, Sup2, Sup3, and Sup4) shown in Fig. 1. The second component is the loss of the final fusion-output saliency map, represented by l_{fuse}. The weights assigned to these loss terms are w_{side}^m and w_{fuse}, respectively.

We compute the loss for each term l using the conventional BCE to calculate the pixel-level comparison between the predicted saliency map and the ground

truth.

$$l = - \sum_{(x,y)}^{(H,W)} [G_{(x,y)} \log P_{(x,y)} + (1 - P_{(x,y)}) \log(1 - P_{(x,y)})],$$

(7)

where (H, W) is the height and width of the image, and (x, y) is the coordinate of a pixel. The ground truth and predicted saliency probability map's pixel values are represented by the symbols $G(x, y)$ and $P(x, y)$, respectively. The goal of the training procedure is to reduce the total loss $\mathcal{L}$ of (6). We select the fusion output l_{fuse} as our final saliency map during the testing process.

IV. EXPERIMENTAL RESULTS

A. Datasets

Training dataset: We train our model using the DUTS-TR dataset, which is a subset of the DUTS dataset [38]. DUTS-TR is curated from the training and validation sets of ImageNet DET [6], and comprises a total of 10 553 images, each with its corresponding ground truth. DUTS is the largest and most widely used dataset for saliency detection. We performed horizontally flipping data augmentation technique, resulting in 21,106 images for training.

Evaluation dataset: We use the following six widely used saliency detection datasets in order to evaluate the detection performances of our model.

ECSSD [36]: The ECSSD (Extended Complex Scene Saliency Dataset) contains semantically significant yet structurally complex and challenging images. This dataset contains 1000 natural images with carefully annotated ground truth saliency masks.

PASCAL-S [23]: This dataset was collected on eight subjects with a 3-second viewing time and the utilization of the Eyelink eye tracker collected from the PASCAL VOC (Visual Object Classes 2010) [9] validation dataset. This dataset contains 850 images featuring multiple salient objects within their scenes, providing a rich and diverse visual context.

HKU-IS [18]: The HKU-IS dataset is a more challenging benchmark for visual saliency detection, aimed at advancing the research and evaluating the performance of visual saliency models. This dataset comprises 4447 challenging images, featuring high-quality pixelwise annotations with characteristics of either low contrast or presence of multiple salient objects.

DUT-OMRON [45]: DUT-OMRON comprises 5168 high-quality nature images meticulously chosen from more than 140 000 images. These images possess dimensions of either $400 \times x$ or $x \times 400$ pixel dimensions, where x is less than 400. Notably, each image features one or more salient objects set against a relatively complex background.

DUTS-TE: DUTS-TE is the test set of the DUTS dataset, which comprises 5019 test images sourced from the ImageNet DET test set and the SUN dataset [40]. This dataset contains highly challenging scenarios for the evaluation of saliency detection models.

SOD [27]: SOD comprises salient object boundaries derived from the Berkeley Segmentation Dataset (BSD) [26]. It consists of 300 particularly challenging images, initially intended for image segmentation.

B. Evaluation Metrics

The probability maps that are produced by deep salient object algorithms often have the same dimension as the input images. In predicted saliency maps, each pixel has a value between 0 and 1 (or [0, 255]). The ground truths are often binary masks, where each pixel is either 0 or 1 (or 0 and 255), with 1 denoting the pixels of the foreground salient object and 0 denoting the background.

To comprehensively evaluate the performance of our model and the quality of the predicted saliency maps against the actual saliency masks, we used the following four evaluating measures: (1) mean absolute error (MAE) [30], (2) maximal F-measure ($maxF\beta$) [2], (3) structure measure (Sm) [10], and (4) enhanced alignment measure (Em) [11]. Using multiple evaluation metrics is crucial when evaluating an SOD model because it provides a more comprehensive assessment of the model's performance across different aspects. Each metric measures a specific quality of the prediction, and no single metric can fully capture all aspects of model generalization. The detailed descriptions of these measures are presented below.

1) F-measure: *F*-measure comprehensively evaluates both precision and recall as:

$$F_\beta = \frac{(1 + \beta_2)\text{Precision} \times \text{Recall}}{\beta_2 \text{Precision} + \text{Recall}}.$$

(8)

Since the rate of recall is not as important as precision, β_2 is empirically set to 0.3 to emphasize precision more.

2) Mean Absolute Error: MAE, or mean absolute error, represents the average difference per pixel between a predicted saliency map and its corresponding ground truth mask. It is used as a metric to accurately assess false negative pixels.

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

(9)

where P and G are the probability map of saliency detection and the corresponding ground truth, respectively, and (H, W) and (x, y) are the (height, width) and the pixel coordinates. A lower MAE value signifies a high degree of similarity between the ground truth and the predicted saliency map.

3) *Structure Measure:* S-measure (Sm) assesses the structural similarity between the predicted saliency map

Dataset	DUT-OMRON				DUTS-TE				ECSSD			
Metrics	MAE	F_β	S_m	E_m	MAE	F_β	S_m	E_m	MAE	F_β	S_m	E_m
DCL	0.132	0.734	0.758	0.763	0.174	0.771	0.776	0.764	0.078	0.910	0.883	0.888
RFCN	0.110	0.742	0.764	0.778	0.09	0.784	0.794	0.839	0.107	0.890	0.852	0.876
UCF	0.132	0.734	0.758	0.763	0.117	0.771	0.776	0.764	0.078	0.910	0.883	0.888
Amulet	0.098	**0.743**	**0.780**	0.784	0.085	0.778	0.802	0.797	**0.059**	**0.912**	**0.893**	**0.911**
SalFAU-Net	**0.080**	0.722	0.755	**0.811**	**0.061**	**0.794**	**0.809**	**0.856**	0.063	0.895	0.862	0.892

The best values are highlighted in bold.

and the binary ground truth. It quantifies how well the predicted salient regions align with the actual salient regions in terms of structure, which are essential for real-world applications. It is defined as the weighted sum of region-aware Sr and object-aware So structural similarity:

$$S = (1 - \alpha)Sr + \alpha So. \tag{10}$$

Typically, α is set to 0.5.

4) Enhanced Alignment Measure: Enhanced alignment measure (Em) incorporates both local pixel matching information and image-level statics by combining local pixel values and the image-level mean or global average value in a single term, which allows for a more comprehensive evaluation of detection algorithms, ensuring that the accuracy and quality of the salient regions are effectively measured.

$$Q_{FM} = \frac{1}{h \times w} \sum_{x=1}^{h} \sum_{y=1}^{w} \phi_{FM}(x, y), \tag{11}$$

where h and w are the height and width of the saliency map, respectively. ϕ_{FM} is enhanced alignment matrix, reflecting the correlation between P and G after subtracting their global means, respectively.

C. Implementation Details

The proposed network is implemented using the PyTorch framework, and training and testing are performed on an NVIDIA GeForce RTX 4070Ti GPU with 12 GB of video memory. The training dataset consists of 10,553 images from the DUTS-TR subset of DUTS [38]. To augment the dataset, each image is horizontally flipped, resulting in a doubled training set with 21,106 images. Prior to feeding the images into the network, they are resized to 320×320 and then cropped to 288 x 288 during training. Model optimization employs the Adam optimizer with default hyperparameter values ($lr = 1e - 3, betas = (0.9, 0.999), eps = 1e - 8, weight\_decay = 0$). The network is trained for approximately 500 000 iterations with a batch size of 12 to ensure convergence of the loss. While testing, the input images are first resized to 320×320 before being inputted into the trained network. The resulting predicted saliency map is then restored to its original dimensions through bilinear interpolation.

D. Comparison With Other Methods

In this section, we evaluate the effectiveness of the proposed model through both qualitative and quantitative analysis. We perform experiments to compare its performance with that of other models, utilizing four evaluation metrics, namely, MAE, F-measure, S-measure, and E-measure. We compare the results of the proposed method with some FCN-based methods, including Amulet [46], DCL [19], RFCN [39], and UCF [47].

1) Quantitative Comparison: The quantitative results on the six evaluation datasets using the four evaluation metrics are reported in Tables I and II. Based

Dataset	HKU-IS				PASCAL-S				SOD			
Metrics	MAE	F_β	S_m	E_m	MAE	F_β	S_m	E_m	MAE	F_β	S_m	E_m
DCL	0.074	0.886	0.866	0.891	0.126	0.824	0.803	0.785	0.164	0.798	0.754	0.755
RFCN	0.089	0.893	0.859	0.906	0.132	0.824	0.798	0.807	0.169	0.797	0.732	0.778
UCF	0.074	0.886	0.866	0.891	0.126	0.824	0.803	0.785	0.164	0.798	0.754	0.755
Amulet	0.052	0.895	0.882	0.912	0.098	**0.833**	**0.819**	0.827	0.0141	**0.802**	**0.759**	**0.791**
SalFAU-Net	**0.044**	**0.896**	**0.885**	**0.927**	**0.091**	0.814	0.801	**0.834**	**0.137**	0.798	0.718	0.759

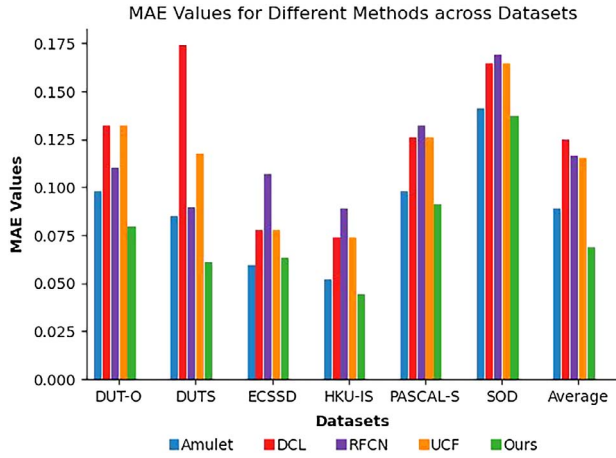The best values are highlighted in bold.

Figure 3. MAE values of Amulet, DCL, RFCN, UCF, and the proposed SalFAU-Net model across six evaluation datasets and the average MAE values of each method across all datasets. A lower MAE value signifies superior performance.

on the results, it is evident that our proposed method outperforms the benchmark methods on DUTS-TE and HKU-IS datasets in all evaluation metrics. On the DUT-OMRON dataset, our model achieves impressive results, outperforming other methods with the best MAE values. Furthermore, on the ECSSD dataset, our model demonstrates competitive performance with the second lowest MAE value of 0.063, surpassed only by Amulet with a slightly lower MAE of 0.059. Furthermore, we calculate the average MAE values for each method across the six datasets. Impressively, our proposed method achieves the lowest average MAE value of 0.068, indicating superior performance compared to other methods. Figure 3 presents the results of the average MAE values of each method, which clearly demonstrate that our model outperforms the comparison algorithms based on the average MAE value.

2) Qualitative Comparison: In addition to quantitative evaluations, we present predicted saliency maps generated by the proposed method and the comparison methods in Fig. 4. The images in the first and second columns of Fig. 4 represent the original input images and their corresponding ground truth saliency maps, respectively. The third column showcases the predicted saliency maps of our proposed method, while the fourth, fifth, sixth, and seventh columns exhibit the results of the comparison methods. The first two rows depict scenarios with multiple salient objects; the third row showcases a single large salient object; the fourth row contains small objects; and the fifth and sixth rows depict images with both small and large salient objects. The last row features relatively low-contrast salient objects. As we can see from Fig. 4, the results demonstrate that SalFAU-Net generates saliency maps more accurately for different challenging scenes, while the comparison methods generate incomplete or noisy saliency maps.

From the qualitative and quantitative results presented above, it is evident that our proposed method

yields competitive results in tackling the challenge of visual saliency detection. These findings also highlight the crucial role of attention mechanisms in enhancing the effectiveness of the visual saliency models, as the proposed model places significant emphasis on extracting highly representative features while effectively eliminating unwanted or noisy features. This emphasis on attention mechanisms not only contributes to the competitiveness of our approach but also enhances the overall performance by prioritizing the extraction of relevant or salient visual information.

E. Failure Cases

The proposed method demonstrated effective SOD in most cases. However, there are some cases where it exhibits limitations. Figure 5 shows some failure cases for the proposed method. In the first column of Fig. 5, the presence of the shadow of the person is erroneously detected as a salient object. This is because the presence of shadows in salient objects may cause a decrease in the visibility or distinguishability of salient objects, making them harder to detect accurately. In the second column, the reflection of the duck is identified as a salient object, which is caused by reflections, which can create distracting image regions, potentially diverting attention away from the true salient objects. The third and fourth columns depict situations where the salient objects have low contrast, causing difficulty for our model in accurate saliency detection. In the last column, although most part of the airplane is detected, the model fails to capture its entirety.

In general, these images are very challenging for most deep learning models to detect accurately. These challenges arise due to the sensitivity of deep learning models to factors such as shadows, reflections, and low contrasts in salient objects. In the future, we will carry out further research aiming to address these problems and develop more accurate saliency detection models.

V. CONCLUSION

In this paper, we proposed SalFAU-Net as an approach for visual saliency detection tasks. Our method integrates an SFM into each decoder block of the Attention U-Net model, which enables efficient generation of saliency maps. The use of an AG module in our method facilitates selective focus on informative regions and suppression of nonsalient regions within an image. The comprehensive evaluation across six diverse SOD datasets, both quantitatively and qualitatively, underscores the effectiveness of our proposed method compared to the benchmark methods. SalFAU-Net not only showcases competitive performance but also highlights the potential of attention-based models in advancing the capabilities of saliency detection models.
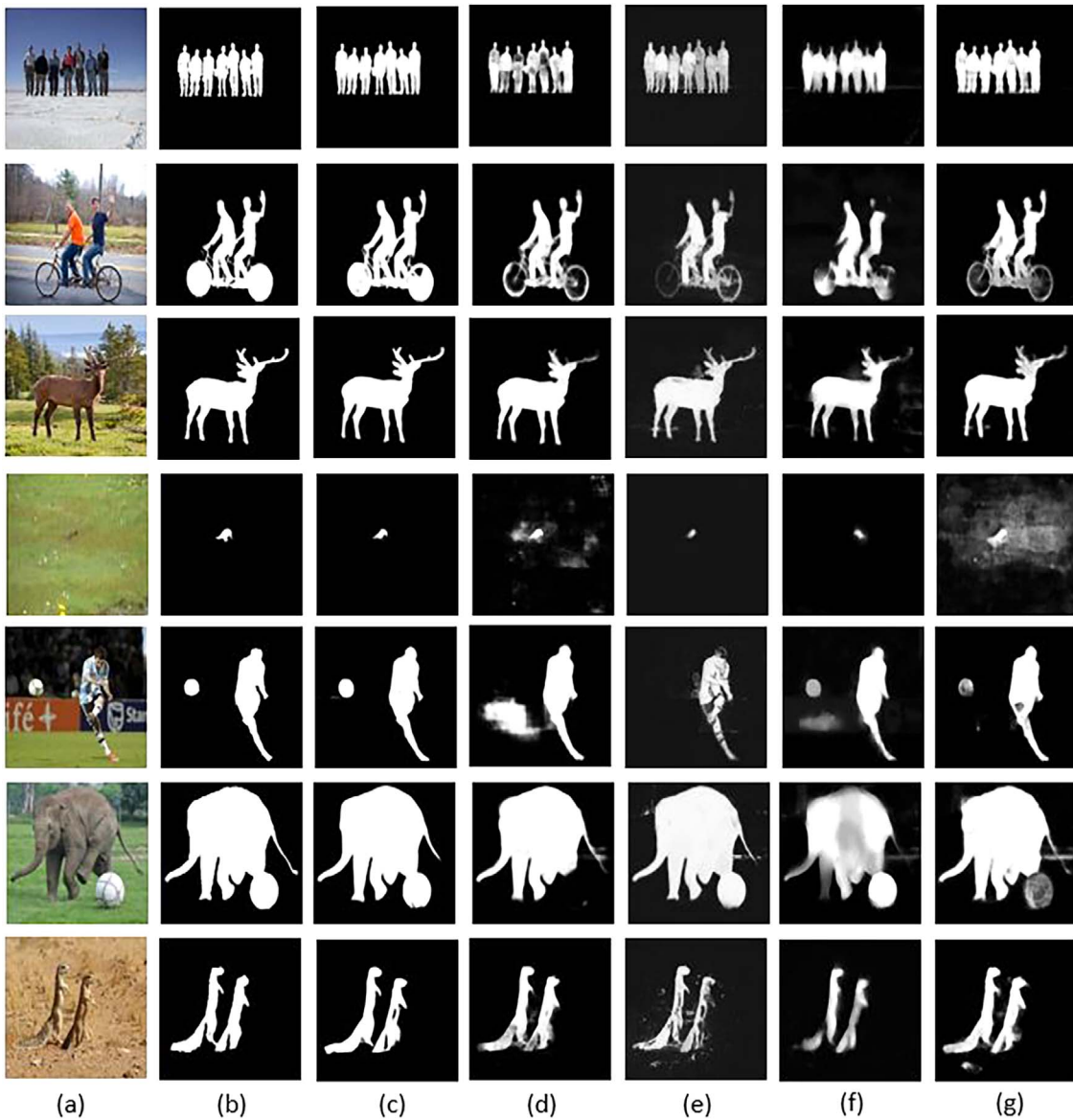
Figure 4.   Qualitative comparison of the proposed method with four other SOTA methods: (a) original image, (b) GT, (c) Ours, (d) Amulet, (e) DCL, (f) RFCN, and (g) UCF.
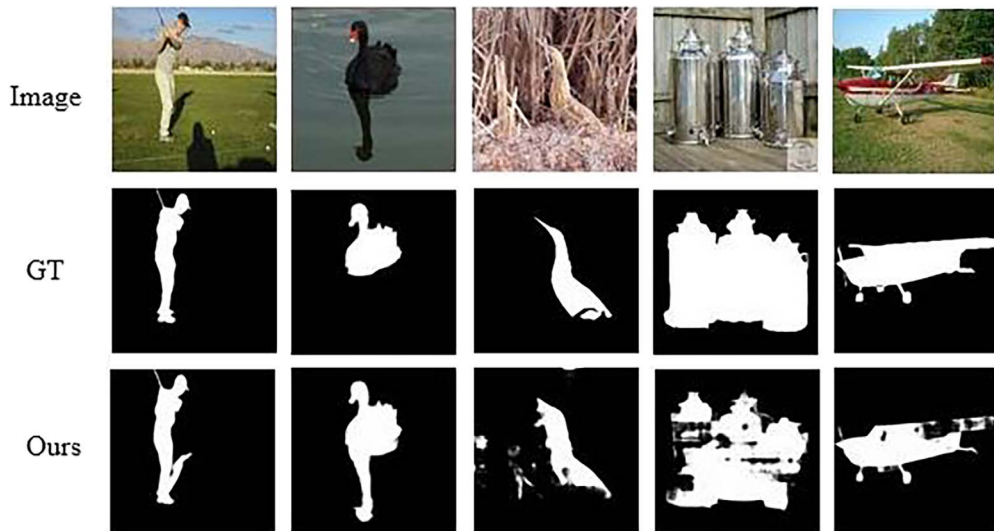

Figure 5.   Failure cases of the proposed method.

REFERENCES

[1] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk
"Salient region detection and segmentation,"
in *Proc. 6th Int. Conf. Comput. Vis. Syst.*, Santorini, Greece,
2008, pp. 66–75.

[2] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk
"Frequency-tuned salient region detection,"
in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009,
pp. 1597–1604.

[3] A. Borji
"What is a salient object? a dataset and a baseline model
for salient object detection,"
IEEE Trans. Image Process., vol. 24, no. 2, pp. 742–756,
Feb. 2015.

[4] A. Borji and L. Itti
"Scene classification with a sparse set of salient regions,"
in *Proc. IEEE Int. Conf. Roboti. Automat.*, 2011,
pp. 1902–1908. IEEE.

[5] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein
"A tutorial on the cross-entropy method,"
Ann. Operations Res., vol. 134, pp. 19–67, 2005.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and L. Fei-Fei
"Imagenet: A. large-scale hierarchical image database,"
in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2009,
pp. 248–255.

[7] M. Donoser, M. Urschler, M. Hirzer, and H. Bischof
"Saliency driven total variation segmentation,"
in *Proc. IEEE 12th Int. Conf. Comput. Vis.* 2009, pp. 817–824.

[8] H. Duan, W. Shen, X. Min, D. Tu, J. Li, and G. Zhai
"Saliency in augmented reality,"
in *Proc. 30th ACM Int. Conf. Multimedia*, 2022,
pp. 6549–6558.

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. M. Winn, and
A. Zisserman
"The pascal visual object classes (VOC) challenge,"
Int. J. Comput. Vis., vol. 88, pp. 303–338, 2010.

[10] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji
"Structure-measure: A. new way to evaluate foreground
maps,"
in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4548–4557.

[11] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji
"Enhanced-alignment measure for binary foreground map
evaluation," 2018, arXiv:1805.10421.

[12] S. H. Fu, D. Xu, and D. Lin
"Object-based multiple foreground segmentation in rgbd
video,"
IEEE Trans. Image Process., vol. 26, no. 3, pp. 1418–1427,
Mar. 2017.

[13] X. Gong, L. Qingge, Q. Liu, and P. Yang
"Improved U-net-like network for visual saliency detection
based on pyramid feature attention,"
Wireless Communi. Mobile Comput., vol. 2022, no. 1,
p. 1108462, 2022.

[14] L. Han, X. Li, and Y. Dong
"Convolutional edge constraint-based U-net for salient
object detection,"
IEEE Access, vol. 7, pp. 48890–48900, 2019.

[15] X. Hu, C.-W. Fu, L. Zhu, T. Wang, and P.-A. Heng
"Sac-net: Spatial attenuation context for salient object
detection,"
IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 3,
pp. 1079–1090, 2020.

[16] Y. Ji, H. Zhang, Z. Zhang, and M. Liu
"Cnn-based encoder-decoder networks for salient object
detection: A comprehensive review and recent advances,"
Inf. Sci., vol. 546, pp. 835–857, 2021.

[17] X. Lei, H. Pan, and X. Huang
"A dilated cnn model for image classification,"
IEEE Access, vol. 7, pp. 124087–124095, 2019.

[18] G. Li and Y. Yu
"Visual saliency based on multiscale deep features,"
in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2015,
2015, pp. 5455–5463 2015.

[19] G. Li and Y. Yu
Deep contrast learning for salient object detection.
in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*,
pp. 478–487, 2016.

[20] H. Li, H. Lu, Zhe Lin, Xiaohui Shen, and Brian Price
"Inner and inter label propagation: Salient object detection
in the wild,"
IEEE Trans. Image Process., vol. 24, no. 10, pp. 3176–3186,
Oct. 2015.

[21] J. Li, Z. Pan, Q. Liu, and Z. Wang
"Stacked U-shape network with channel-wise attention for
salient object detection"
IEEE Trans. Multimedia, vol. 23, pp. 1397–1409, 2020.

[22] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang
Saliency detection via dense and sparse reconstruction.
in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2976–2983.

[23] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille
"The secrets of salient object segmentation,"
in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2014,
pp. 280–287.

[24] X. Liu, L. Song, S. Liu, and Y. Zhang
Sustainability, vol. 13 no. 3, 2021, Art. no. 1224.

[25] J. Long, E. Shelhamer, and T. Darrell
"Fully convolutional networks for semantic segmentation,"
in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015,
pp. 3431–3440.

[26] D. Martin, C. Fowlkes, D. Tal, and J. Malik
"A database of human segmented natural images and its
application to evaluating segmentation algorithms and
measuring ecological statistics,"
in *Proc. 8th IEEE Int. Conf. Comput. Vis.*, 2001, pp. 416–423.

[27] V. Movahedi and J. H. Elder
"Design and perceptual validation of performance
measures for salient object segmentation,"
in *Proc. IEEE Comput. Soci. Conf. Comput. Vis. Pattern
Recognit. Workshops*, 2010, pp. 49–56.

[28] O. Oktay, J. Schlemper, L.L. Folgoc, M. Lee, M. Heinrich,
K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla,
B. Kainz et al.
"Attention U-net: Learning where to look for the pancreas,"
2018, arXiv:1804.03999.

[29] A. Pal, S. Mondal, and H. I. Christensen
"Looking at the right stuff"-guided semantic-gaze for
autonomous driving.
in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*,
2020, pp. 11883–11892.

[30] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung
"Saliency filters: Contrast based filtering for salient region
detection,"
in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012,
pp. 733–740.

[31] C. Qin, G. Zhang, Y. Zhou, W. Tao, and Zhiguo Cao
"Integration of the saliency-based seed extraction and
random walks for image segmentation,"
Neurocomputing, vol. 129, pp. 378–391, May 2014.

[32] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and
M. Jagersand
"U2-net: Going deeper with nested U-structure for salient
object detection,"
Pattern Recognit., vol. 106, 2020, Art. no. 107404.

[33] Z. Ren, S. Gao, L.-T. Chia, and I.W.-H. Tsang
"Region-based saliency detection and its application in
object recognition,"
IEEE Trans. Circ. Syst. Video Tech, vol. 24, no. 5, pp. 769–779,

May 2014.

[34] O. Ronneberger, P. Fischer, and T. Brox
U-net: Convolutional networks for biomedical image segmentation.
in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, Munich, Germany, 2015, pp. 234–241.

[35] U. Rutishauser, D. Walther, C. Koch, and P. Perona
"Is bottom-up attention useful for object recognition?,"
in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2004, pp. II–II.

[36] J. Shi, Q. Yan, L. Xu, and J. Jia
"Hierarchical image saliency detection on extended CSSD,"
IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 4, pp. 717–729, Apr. 2016.

[37] I. Ullah, M. Jian, S. Hussain, J. Guo, H. Yu, X. Wang, and Y. Yin
"A brief survey of visual saliency detection,"
Multimed. Tools Appl., vol. 79, pp. 34605–34645, 2020.

[38] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan
"Learning to detect salient objects with image-level supervision,"
in *Proc. IEEE Conf. Computer Vis. Pattern Recognit.*, 2017, pp. 3796–3805.

[39] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan
"Saliency detection with recurrent fully convolutional networks,"
in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 825–841. Springer, 2016.

[40] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba
"Sun database: Large-scale scene recognition from abbey to zoo,"
in *Proc. Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3485–3492.

[41] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han
"Oriented r-cnn for object detection,"
in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3520–3529.

[42] Y. Xie, H. Lu, and M.-Hsuan Yang
"Bayesian saliency via low and mid level cues,"
IEEE Trans. Image Process., vol. 22, no. 5, pp. 1689–1698, May 2013.

[43] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio
"Show, attend and tell: Neural image caption generation with visual attention,"
in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015 pp. 2048–2057.

[44] Q. Yan, L. Xu, J. Shi, and J. Jia
"Hierarchical saliency detection,"
in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1155–1162.

[45] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang
"Saliency detection via graph-based manifold ranking,"
in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3166–3173.

[46] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan
"Amulet: Aggregating multi-level convolutional features for salient object detection,"
in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 202–211.

[47] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin
"Learning uncertain convolutional features for accurate saliency detection",
in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 212–221.

[48] Q. Zhang, Y. Shi, and X. Zhang
"Attention and boundary guided salient object detection,"
Pattern Recognit., vol. 107, pp. 107484, 2020.

[49] T. Zhao and X. Wu
"Pyramid feature attention network for saliency detection,"
in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3085–3094.

Kassaw Abraham Mulat received his bachelor's degree in software engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2021. He is currently pursuing his master's degree at the School of Computer Science, China West Normal University, Nanchong, China. His research interests include computer vision, pattern recognition, salient object detection, and medical image processing.

Zhengyong Feng received the B.S. degree in Physics from China West Normal University, Nanchong, China, in 2001, the M.S. degree in Communication and Information Systems from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2004, and the Ph.D. degree in Communication and Information Systems from University of Electronic Science and Technology of China, Chengdu, China, in 2013. From 2010 to 2011, he was a visiting PhD student at the School of Computer Science and Engineering, University at Buffalo, the State University of New York, USA. From 2004 to 2007, he was a Research Assistant at the School of Physics and Electronic Information, China West Normal University. Since 2017, he has been a Professor at the School of Electronic Information Engineering, China West Normal University. He is the author of more than 80 articles. His research interests include machine vision with deep learning, optimal control with MPC (Model Predictive Control), and RL (Reinforcement Learning).

Tegegne Solomon Eshetie received his bachelor's degree in Electronic Information Engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2021. He is currently pursuing his master's degree at the School of Computer Science, China West Normal University, Nanchong, China. His research interests include computer vision, medical image processing machine learning, and deep learning.

Ahmed Endris Hasen received his Bachelor's and Master's degree in software engineering from the University of Electronic Science and Technology of China (UESTC) in 2021 and 2023, respectively. Currently, he is a PhD Student at Paul C. Lauterbur Research Center, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences. His research interests include Deep Learning, Computer vision, Image processing, and medical imaging.