Enhanced U-Net Models with Hybrid Attention Elements for Medical Image Segmentation

Piyushkumar Banugariya
Department of Computer Science
Royal Holloway, University of London
Surrey, TW20 0EX, UK
piyushkumar.banugariya.2023@live.rhul.ac.uk

Li Zhang
Department of Computer Science
Royal Holloway, University of London
Surrey, TW20 0EX, UK
li.zhang@rhul.ac.uk

Yonghong Yu
College of Tongda
Nanjing University of Posts and Telecommunications
Nanjing, 210049, China
yuyh@njupt.edu.cn

Vivian Sedov
Department of Computer Science
Royal Holloway, University of London
Surrey, TW20 0EX, UK
vivian.sedov.2020@live.rhul.ac.uk

2025 International Joint Conference on Neural Networks (IJCNN) | 979-8-3315-1042-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/IJCNN64981.2025.11228708

***Abstract*—This research presents an enhanced deep learning-based approach for medical image segmentation by integrating multiple attention mechanisms into the U-Net architecture. Specifically, the proposed models incorporate seven advanced attention mechanisms, including Convolutional Block Attention Module, Attention Gate, Squeeze-and-Excitation, Halo, Coordinate, Spatial, and Triplet Attention strategies, to improve the model's capability to capture both channel and spatial contextual information within medical images. These attention-enhanced U-Net variants are rigorously evaluated on multiple medical imaging datasets, demonstrating significant improvements in segmentation accuracy compared to benchmark models such as U-Net and DeepLabV3+. The results indicate superior performance, especially in complex scenarios with overlapping structures and fine organ/lesion details, showcasing the effectiveness of attention mechanisms for improving segmentation in medical imaging.**

Keywords—Image Segmentation, Deep Learning, Attention Mechanism, Hybrid Attention Models, Medical Imaging.

I. Introduction

Image segmentation plays a crucial role in medical imaging by enabling the accurate identification and delineation of structures of interest, such as organs, tissues, or pathological regions. Segmentation is fundamental in applications like disease diagnosis, treatment planning, and health monitoring. The precision and reliability of image segmentation directly influence the outcomes of critical procedures, such as tumor detection, organ delineation, and the tracking of disease progression.

Traditionally, Convolutional Neural Networks (CNNs) such as U-Net and DeepLabV3+ have demonstrated significant success in segmenting medical images due to their ability to learn hierarchical features. However, these models often face challenges when handling complex scenes that require capturing long-range dependencies and intricate spatial relationships. When objects are closely packed, exhibit complex boundaries, or appear amidst varied backgrounds, conventional CNN architectures may struggle to accurately preserve boundaries and contextual information. This limitation becomes especially evident in medical imaging, where precise segmentation of fine structures is often necessary for making accurate clinical decisions.

Recent advancements in deep learning have shown that attention mechanisms can significantly enhance the performance of CNN-based architectures by allowing models to dynamically focus on the most relevant features. Attention mechanisms such as Convolutional Block Attention Module (CBAM), Squeeze-and-Excitation (SE), Coordinate, and Halo Attention mechanisms and others have shown great promise in improving the capacity of models to capture both local and global features simultaneously. By integrating these mechanisms into traditional CNN models, it is possible to overcome the inherent limitations of purely convolutional approaches.

This study aims to develop an enhanced U-Net architecture by integrating multiple attention mechanisms to improve medical image segmentation performance. Specifically, we have incorporated seven attention mechanisms, i.e. CBAM, Attention Gate, SE Attention, Halo Attention, Coordinate Attention, Triplet Attention, and Spatial Attention, at strategic points within the U-Net architecture. These mechanisms allow the model to adaptively emphasize relevant features, thus improving its ability to capture complex relationships within medical images.

The novel aspects of this research are elaborated as follows.

- We present seven attention-enhanced U-Net models for medical image segmentation, i.e. UNet integrated with Attention Gate, CBAM, SE, Halo, Coordinate, Triplet, Spatial attention mechanisms, respectively, each tailored to improve the segmentation accuracy for various types of medical images.
- Seven medical imaging datasets are used to evaluate these attention-enhanced models, including Kvasir-SEG, ISIC 2017, Data Science Bowl, CVC-ClinicDB, Kvasir Instrument, HyperKvasir and Retina Blood Vessel Segmentation datasets, for gastrointestinal polyp, skin lesion, medical instrument, and retina vessel segmentation problems. Our results demonstrate significant improvements in segmentation accuracy compared to those of traditional U-Net and other benchmark models. These empirical studies highlight practical insights into the application of attention mechanisms in image segmentation, showcasing their

advantages in enhancing performance in scenarios where precision is of utmost importance.

II. Related Work

Recent advancements in medical image segmentation have introduced a variety of methodologies to improve segmentation performance, particularly through the integration of attention mechanisms and enhanced architectures. This section discusses key studies in the field, focusing on their methodologies, novel contributions, and evaluation results.

Attention mechanisms have become a cornerstone in improving deep learning models for medical image segmentation. One such approach involves using attention modules to refine the feature extraction process and improve segmentation accuracy. The **SE Attention** method [1] introduced a novel way to recalibrate channel-wise feature responses by learning the importance of each feature map through global pooling and feature reweighting. This technique allows the model to focus on the most informative parts of the image, leading to improved segmentation accuracy. SE networks were shown to outperform baseline models on tasks involving complex structures, such as organ and tumour segmentation, by emphasizing critical features while suppressing irrelevant information [1]. The use of SE modules has proven particularly effective in tasks where subtle differences in texture and boundary details significantly impact performance.

Another noteworthy development is the **CBAM** [2], which incorporates both channel and spatial attention mechanisms into deep neural networks. CBAM enhances the model's ability to identify important spatial regions and feature channels, improving segmentation outcomes for complex medical images, such as those involving elongated structures like blood vessels. By applying attention sequentially to channel and spatial dimensions, CBAM helps models better capture and refine the most relevant features in medical image segmentation tasks. Its resulting networks outperformed traditional CNNs by focusing on the most critical parts of the image, achieving higher accuracy in challenging datasets [2]. CBAM's effectiveness has been demonstrated in various segmentation tasks, particularly where identifying fine structures is crucial.

The combination of attention mechanisms with U-Net has also been a focus of many studies. **Attention U-Net** [3] introduced **Attention Gates** into the U-Net architecture, allowing the model to selectively focus on relevant features while ignoring background noise. This improvement was particularly beneficial in segmenting organs and lesions, where the boundaries are often difficult to discern. By integrating Attention Gates into the encoder-decoder architecture, Attention U-Net significantly improved the segmentation performance over the baseline U-Net, making it especially useful for tasks that involved identifying small or irregularly shaped objects in medical images [3]. The ability of Attention Gates to suppress irrelevant regions helps ensure that the model focuses on the critical parts of the image, resulting in better overall segmentation outcomes.

In addition to attention modules, novel architectures such as **U-Net++** [4] have further improved the U-Net framework by introducing nested skip connections and dense convolutional blocks. The key innovation of U-Net++ lies in its ability to improve feature representation across multiple scales, which is essential for medical image segmentation tasks where objects vary greatly in size. This architectural enhancement allows U-Net++ to capture both coarse and refined details more effectively than the standard U-Net. Evaluations have shown that U-Net++ achieves better segmentation performance than its predecessors, especially in tasks requiring multi-scale feature extraction [4]. This demonstrates the importance of refining the architecture to improve feature fusion and representation in medical image segmentation tasks.

Coordinate Attention [5] is another attention mechanism that improves spatial attention by embedding positional information into the feature representation process. Unlike traditional attention mechanisms that discard spatial relationships during global pooling, Coordinate Attention preserves the positional dependencies between different regions of the image, which is particularly beneficial for segmenting elongated structures such as blood vessels. This method has proven effective in tasks such as retinal blood vessel segmentation, where spatial dependencies are crucial for accurate segmentation [5]. The introduction of positional information ensures that the model captures not only feature importance but also spatial relationships, leading to improved segmentation accuracy.

Finally, transformer-based architectures, such as **TransUNet** [6], have introduced the concept of global self-attention to medical image segmentation. TransUNet combines the strengths of CNN-based local feature extraction with transformers' global attention, allowing the model to capture both short- and long-range dependencies in medical images. This hybrid approach has proven particularly useful in tasks such as multi-organ segmentation, where the ability to understand both local details and global context is critical. The key advantage of TransUNet is its ability to outperform traditional CNN-based architectures, particularly in complex tasks that involve segmenting multiple anatomical structures with varying shapes and sizes [6]. This highlights the importance of global attention in improving segmentation performance.

III. The Proposed Methodology

In this section, we detail the methodology employed to enhance the U-Net architecture with attention mechanisms for improving precision in medical image segmentation. Our approach involves integrating seven attention modules—such as Attention Gate, SE, CBAM, Coordinate, Halo, Triplet and Spatial Attention strategies—into the U-Net architecture to address limitations in capturing both local and global contextual features. By utilizing attention mechanisms at strategic points within the network, we aim to significantly improve the segmentation performance, especially in complex scenarios where medical images exhibit overlapping structures and clutter organs/lesions along with medical instruments, with complex boundary details.

A. *Integration of Attention Mechanisms in U-Net*

The U-Net architecture is a popular choice for medical image segmentation due to its symmetric encoder-decoder structure. It comprises an encoder with down-sampling convolutional layers, for hierarchical feature learning, a decoder with upsampling operations through transposed convolutions, as well as residual connections facilitating the preservation of fine-grained details.

While U-Net is effective in many cases, it struggles to capture long-range dependencies and contextual relationships, which are essential for segmenting complex medical images. This shortcoming becomes particularly evident when dealing with highly heterogeneous structures such as tumours or blood vessels, where precise boundary delineation is crucial.

To enhance the performance of the U-Net architecture, we integrate attention mechanisms into the skip connections in U-Net. Figure 1 shows the system architecture. These skip connections allow for the transfer of high-resolution features from the encoder to the decoder, facilitating the recovery of fine details during upsampling. By incorporating attention mechanisms into these connections, we enable the network to selectively focus on the most relevant features, thus improving segmentation performance. Below, we describe the different attention modules integrated into the skip connections in the proposed enhanced U-Net architectures.

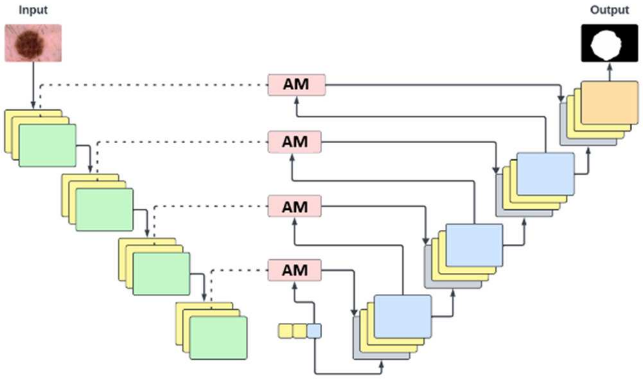

Fig. 1. The proposed U-Net variant architecture integrating diverse attention strategies in the skip connections (where AM represents each of seven adopted attention mechanisms).

Symbols Explained:

- $\boldsymbol{F}$: Input feature map defined with height H, width W, and number of channels C, respectively.
- x: Squeezed channel descriptor vector used to represent global information across spatial dimensions. s: Excitation vector that recalibrates the feature map by scaling channel responses. α : Attention coefficient used to filter the feature map, enhancing relevant features.
- σ: Sigmoid activation function, used to normalize the attention weights between 0 and 1.
- $q,\ k, and\ v$: Query, key, and value matrices used in self-attention mechanisms like Halo Attention.

1) *The SE Module*

In U-Net, feature maps are passed between the encoder and decoder via skip connections. One limitation of the standard U-Net architecture is that it treats all channels equally, without considering their relative importance. By integrating **SE** blocks into the skip connections or convolutional layers of U-Net, the network can learn which channels are most relevant for the segmentation task.

Specifically, the SE module [1] is designed to enhance the representational power of CNNs by recalibrating channel-wise feature responses. It consists of two main operations, i.e. squeeze and excitation.

Squeeze: The squeeze operation compresses the spatial dimensions of the feature maps into a single channel-wise descriptor. This is achieved through global average pooling, which collapses the spatial dimensions $H \times W$ into a scalar value for each channel. Equation (1) defines the detailed operation [1].

$$x_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{F}(i,j,c) \qquad (1)$$

where $x_c \in R^C$ represents the squeezed vector for channel c.

Excitation: The excitation operation takes the squeezed channel descriptor x and passes it through a pair of dense layers. The first layer reduces the dimensionality of the descriptor by a reduction ratio r, and the second layer increases it back to the original size. This allows the model to learn a set of weights for each channel. The output is then activated using the sigmoid function, and these weights are used to recalibrate the feature maps [1]. Equation (2) shows the respective process.

$$s = \sigma\big(W_2 \cdot \mathrm{ReLU}(W_1 \cdot x)\big) \qquad (2)$$

Recalibration: The recalibration step applies the learned excitation weights s to F via channel-wise multiplication, to emphasize the most important channel detals, as indicated in Equation (3).

$$\hat{F_c} = s_c \cdot F_c \qquad (3)$$

As indicated in Figure 1, the SE attention is integrated within the skip connections in different network stages to bridge encoder feature representations with decoder feature learning while extracting channel dependencies from the encoder feature maps.

2) *The CBAM Module*

The **CBAM** method [2] improves feature learning by sequentially applying attention mechanisms along both the channel and spatial dimensions. This allows the network to focus on the most important channels and spatial regions in the feature maps. Similar to the SE module, the CBAM attention method is also integrated within the skip connections at different network stages as indicated in Figure 1 to improve spatial and channel context extraction.

Channel Attention: CBAM first applies channel attention, which highlights the most important channels by using global average and max pooling operations followed by multi-layer perceptrons (MLPs) to compute the channel-wise attention feature map, $S_c(F)$, as depicted in Equation (4).

$$S_c(F) = \sigma\Big(MLP\big(GAP(F)\big) + MLP\big(GMP(F)\big)\Big) \qquad (4)$$

Spatial Attention: After channel attention, spatial attention is applied to highlight the most important spatial regions in the feature map using pooling across channels followed by a convolution, as indicated in Equation (5).

$$S_s(F) = \sigma\big(f^{7\times 7}([\mathrm{GAP}(F);\mathrm{GMP}(F)])\big) \qquad (5)$$

where $f^{7\times 7}$ is a 7x7 convolution.

Integrating multiple **CBAM** elements into U-Net enhances its ability to focus on both the most relevant channels and significant spatial regions, which is crucial for accurate segmentation. By first applying channel attention, CBAM helps U-Net emphasize the most informative feature maps, improving its ability to detect critical structures in medical images. The subsequent spatial attention refines the focus on specific regions of interest, such as organs or lesions, improving localization accuracy. This dual attention mechanism enhances U-Net's overall performance without adding significant computational overhead, making it effective for tasks that require precise boundary detection with lightweight computational costs.

3) Coordinate Attention

Besides applying SE and CBAM, we also integrate Coordinate Attention in the skip connection as shown in Figure 1 to enhance channel-wise dependency extraction. Specifically, **Coordinate Attention** [5] improves the standard channel attention mechanism by embedding positional information directly into the attention process. Unlike traditional attention mechanisms that discard positional information during global pooling, Coordinate Attention preserves spatial relationships by factorizing attention into two 1D feature encoding processes along the height and width axes.

Coordinate Information Embedding: Instead of applying global pooling across both height H and width W simultaneously, Coordinate Attention applies pooling along each axis separately, producing two directional attention maps, as shown in Equation (6). This retains spatial information while allowing the network to capture cross-channel dependencies.

$$x_h(i,:) = \frac{1}{H}\sum_{i=1}^{H} F(i,:,:), \quad x_w(:,j) = \frac{1}{W}\sum_{j=1}^{W} F(:,j,:) \tag{6}$$

Recalibration: The directional attention maps are used to generate attention weights, which are applied to recalibrate the feature maps based on both channel importance and positional relevance.

As shown in Figure 1, integrating multiple **Coordinate Attention** methods into U-Net enhances segmentation performance by embedding positional information directly into the attention process. This allows U-Net to capture long-range dependencies and spatial relationships more effectively, improving its ability to segment objects with complex shapes and boundaries, as evidenced in our experimental studies. By preserving spatial information during attention calculation, Coordinate Attention enables U-Net to better localize structures in medical images, such as organs or lesions. Additionally, the lightweight design of Coordinate Attention ensures that these improvements are achieved without significantly increasing the model's complexity, making it suitable for real-time medical segmentation tasks.

4) Attention Gate

Attention Gate (AG) [3] is also embedded in the skip attention as showcased in Figure 1 to enhance both spatial and semantic rich feature representations from both encoder and decoder networks in U-Net in our studies. Precisely, AG is a mechanism that enables networks to focus on the most relevant features by selectively filtering out irrelevant or noisy information. It works by learning spatial attention maps that highlight important features in the input data, which is particularly useful in medical image segmentation tasks where certain regions of interest (e.g., tumours or organs) need to be emphasized. The AG method takes two inputs: the feature map from the encoder and a gating signal from the decoder. The attention map is computed by comparing these two inputs, and a soft attention coefficient is generated to weight the encoder features. This allows only the most relevant information to be passed through the skip connections to the decoder, filtering out irrelevant details.

Integrating multiple AG elements into U-Net enhances the segmentation performance by allowing the model to focus on the most important regions of the image. In U-Net, skip connections carry information from the encoder to the decoder. By adding Attention Gates, these connections are refined, and only relevant features are passed to the decoder. This selective filtering improves the model's ability to segment important structures, such as lesions or organs, by focusing on areas that matter the most while reducing the influence of irrelevant background information. As observed in our experiments, Attention Gates help improve boundary precision and object localization in medical images, contributing to more accurate and efficient segmentation outcomes without adding significant computational cost.

5) Halo Attention

Halo Attention [7] is another attention strategy adopted in the skip connections in the U-Net as the cases for the aforementioned attention methods. Specifically, Halo Attention is a type of local self-attention mechanism designed to reduce the computational complexity of global self-attention. Instead of applying attention across the entire feature map, Halo Attention operates within non-overlapping local windows (or "halos") while also allowing information exchange between neighbouring windows. This approach captures both local and global contexts without the computational overhead associated with traditional self-attention methods. Halo Attention splits the feature map into small, local windows. Within each window, attention is computed to capture local context. The windows are padded (the "halo" region) to allow cross-window interaction, ensuring that the model can capture long-range dependencies without needing global attention across the entire image.

As indicated in our empirical studies, integrating multiple **Halo Attention** methods into U-Net improves its ability to capture both local and global context in medical images while maintaining computational efficiency. Traditional self-attention mechanisms are expensive, especially for high-resolution images, but Halo Attention reduces this complexity by focusing on local regions while still allowing some global context through the halo overlap. This is particularly useful in segmentation tasks where both fine details and broader spatial relationships matter. By applying Halo Attention in U-Net, the model can effectively capture intricate details, such as boundaries of organs or lesions, while maintaining scalability for larger images. This leads to better segmentation accuracy, especially in complex medical images where both global structure and local details are important.

6) Triplet Attention

As an advanced attention mechanism designed to capture dependencies across three dimensions—spatial, channel, and temporal (or depth), **Triplet Attention** [8] is also integrated in the skip connections as shown in Figure 1. It simultaneously processes spatial relationships, recalibrates channel-wise feature maps, and captures temporal or depth-based dependencies, making it particularly suited for tasks involving 3D medical images, such as MRI or CT scans. This multi-dimensional approach allows the network to model complex interactions between the three aspects, potentially leading to improved segmentation in volumetric or sequential data.

Channel Attention: As defined in Equation (4), the channel attention mechanism embedded in Triplet Attention is used to extract inter-channel dependencies effectively. **Spatial Attention:** Next, the spatial attention strategy defined in Equation (5) is applied to highlight the most important regions in the image. **Temporal/Depth Attention:** Finally, attention is applied across the depth or temporal dimension (for 3D images or time-series data). As shown in Equation (7), temporal attention focuses on learning dependencies between different slices/crops of the image:

$$S_t(F) = \sigma(W_t F) \tag{7}$$

where W_t is a weight matrix that captures the inter-dependencies between the slices or time steps.

As showcased in our experiments, integrating multiple **Triplet Attention** elements into U-Net allows the model to simultaneously capture important spatial regions, recalibrate channels, and process depth information, which has improved segmentation performance for 3D or sequential medical data. This multi-faceted attention helps the network focus on critical structures and features across all dimensions, making it a powerful addition for our volumetric image segmentation tasks.

7) Spatial Attention

As illustrated in Figure 1, we have incorporated multiple Spatial Attention elements in the skip connections in different network stages to better extract inter-spatial relationships. Precisely, as defined in Equation (5), **Spatial Attention** [2] aims to highlight **where** in the image the most relevant features are located, helping to refine localization and segmentation tasks.

As evidenced in our empirical studies, incorporating multiple **Spatial Attention** methods into U-Net enhances the model's ability to focus on specific regions of interest, making it highly effective for tasks that require precise localization. By applying attention across the spatial dimensions, U-Net can better detect and refine the boundaries of organs, or other critical structures in medical images. This is especially useful in segmentation tasks where fine details, such as edges or boundaries, are crucial for accurate pixel-wise classification.

IV. Evaluation

In this section, we report the experimental results of our attention-enhanced U-Net models across various medical imaging datasets. The empirical studies of the proposed U-Net variants integrated with diverse attention methods with respect to each dataset are conducted. We also compare our models against baseline systems and provide insights into how attention mechanisms improve performance.

A. Datasets

Seven public medical imaging datasets covering a variety of segmentation tasks are adopted to test model efficiency.

(1) Kvasir-SEG Dataset: This dataset contains 1,000 images of gastrointestinal polyps captured during endoscopic examinations. It is critical for training models in polyp detection, which is essential in colonoscopy procedures for identifying precancerous lesions. (2) **ISIC 2017 (International Skin Imaging Collaboration):** This dataset consists of 2,594 images of skin lesions and is used to train models for melanoma detection. The dataset presents challenges due to the significant variability in lesion size, shape, and colour. (3) **Data Science Bowl Dataset:** Comprising 670 images of cellular nuclei captured via microscopy, this dataset is fundamental for tasks like nuclei segmentation for cancer diagnosis and other biological studies. (4) **CVC-ClinicDB Dataset:** This dataset includes 612 images of colorectal polyps extracted from colonoscopy videos, enhancing the model's ability to perform segmentation in dynamic, video-based data, which is essential for real-time applications. (5) **Kvasir Instrument Dataset:** Containing 590 images of medical instruments seen in endoscopic images, this dataset is used to train models to accurately differentiate between medical tools and human tissues, since accurate segmentation of instruments is crucial during surgical procedures. (6) **HyperKvasir Dataset:** A dataset containing 1,000 images representing various gastrointestinal conditions. It is used to ensure that the model can generalize across different gastrointestinal diseases. (7) **Retina Blood Vessel Segmentation Dataset:** This dataset contains 100 retinal fundus images used for segmenting blood vessels. Precise segmentation of retinal vessels is critical for diagnosing ocular diseases like diabetic retinopathy.

B. Experimental Setup and Evaluation Criteria

To evaluate the performance of each network, we used three key metrics: Dice score, Accuracy, and Mean Intersection over Union (mIoU). These metrics were selected to provide a balanced understanding of how well the models performed in terms of pixel-level accuracy and overall segmentation quality.

C. Results for Each Dataset

This section provides the comprehensive analysis of the performance of various U-Net models enhanced with different attention mechanisms across multiple datasets. The primary objective is to evaluate how each attention module influences the segmentation accuracy and other relevant metrics on diverse types of medical imaging data.

1) Kvasir-SEG Dataset

We present the evaluation results for each attention U-Net model with respect to the Kvasir-SEG dataset in Table I.

For the **Kvasir-SEG** dataset, which contains 1,000 polyp images from gastrointestinal endoscopy procedures, the baseline U-Net achieved 0.6056 (Dice) and 0.5010 (mIoU). Both **U-Net + SE** and **U-Net + CBAM** significantly outperformed the baseline. **U-Net + SE** achieved the best scores, **0.7968 (Dice)** and **0.7016 (mIoU)**, likely due to the SE module's ability to recalibrate channel-wise feature responses, helping the model to focus on relevant polyp structures and suppress less important information. This is crucial for the detection of small and irregular polyps, where capturing boundaries accurately is key. In addition, **U-Net + CBAM** achieved the second best resuslts, **0.7346 (Dice)** and **0.6276 (mIoU)**, by applying attention across both channel and spatial dimensions. This dual attention allowed the model to better localize polyps by refining spatial relationships while maintaining channel-wise importance. The combination of these attention mechanisms helped the model better identify polyp boundaries and led to more accurate and refined segmentation outcomes compared to those of the baseline U-Net.

TABLE I. RESULTS OF THE KVASIR-SEG DATASET

Model	*Dice*	*Accuracy*	*mIoU*
U-Net Baseline	0.6056	0.91	0.501
U-Net + SE	**0.7968**	**0.9443**	**0.7016**
U-Net + CBAM	0.7346	0.928	0.6276
U-Net + Coord Attention	0.4455	0.885	0.3417
U-Net + Attention Gate	0.6113	0.8959	0.4983
U-Net + Halo Attention	0.38	0.8708	0.2766
U-Net + Triplet Attention	0.6709	0.9184	0.5616
U-Net + Spatial Attention	0.6946	0.9242	0.5837

2) Data Science Bowl Dataset

We present evaluation results for each enhanced U-Net model with respect to the Data Science Bowl dataset in Table II.

For the **Data Science Bowl** dataset, which consists of 670 microscopy images for cellular nuclei segmentation, the baseline U-Net achieved 0.7797 (Dice) and 0.6784 (mIoU). U-Net variants integrated with SE, CBAM, Coordinate, Halo, Triplet and Attention Gate methods, respectively, depict significant improvements.

TABLE II. RESULTS OF THE DATA SCIENCE BOWL DATASET

Model	*Dice*	*Accuracy*	*mIoU*
U-Net Baseline	0.7797	0.9208	0.6784
U-Net + SE	0.8035	0.9434	0.6943
U-Net + CBAM	0.8118	0.9322	0.7175
U-Net + Coord Attention	0.8158	0.9422	0.7105
U-Net + Attention Gate	**0.8325**	**0.9378**	**0.7325**
U-Net + Halo Attention	0.8136	0.939	0.7032
U-Net + Triplet Attention	0.8105	0.9425	0.7038
U-Net + Spatial Attention	0.7693	0.9312	0.6554

In particular, **U-Net + Attention Gate** achieved the best results, **0.8325 (Dice)** and **0.7325 (mIoU)**, benefiting from Attenion Gate's ability to recalibrate feature importance from a global perspective, which is particularly useful in enhancing nuclei segmentation by focusing on the critical structures and suppressing less relevant channels. **U-Net + CBAM** achieved the second best segmentation performance, **0.8118 (Dice)** and **0.7175 (mIoU)**. The combination of channel and spatial attention in CBAM allowed the model to better capture the boundaries of the nuclei, improving both localization and feature selection. This led to more accurate and detailed segmentation compared to the baseline U-Net, as the model is able to highlight the most relevant nuclei structures while ignoring background noise.

3) CVC-ClinicDB Dataset

The effectiveness of each attention U-Net is also demonstrated in Table III for evaluating CVC-ClinicDB dataset.

For the **CVC-ClinicDB** dataset, which consists of 612 colorectal polyp images from colonoscopy videos, the baseline U-Net achieved 0.4933 (Dice) and 0.3976 (mIoU). U-Net + SE, U-Net + Triplet Attention, and U-Net + Corrdinate Attention significantly improved segmentation performance against other attention networks. **U-Net + Triplet Attention** achieved the best performance in this dataset, with **0.8070 (Dice)** and **0.7226 (mIoU)**. Triplet Attention captures dependencies across spatial, channel, and depth dimensions, which is especially useful for video-based data where both local and long-range dependencies are important for tracking polyps across multiple frames. This comprehensive attention mechanism helped the model improve its ability to segment polyps in real-time applications, resulting in substantial performance gains over the baseline U-Net. **U-Net + Coordinate Attention** obtains the second best performance with **0.7324 (Dice)** and **0.64 (mIoU)**. The Corrodinate Attention module recalibrates the channel-wise feature responses with position information, allowing the model to better focus on polyp structures across video frames, resulting in more accurate detection and boundary delineation.

TABLE III. RESULTS OF THE CVC-CLINICDB DATASET

Model	*Dice*	*Accuracy*	*mIoU*
U-Net Baseline	0.4933	0.9568	0.3976
U-Net + SE	0.72	0.9714	0.6158
U-Net + CBAM	0.5023	0.9581	0.4047
U-Net + Coord Attention	0.7324	0.9747	0.64
U-Net + Attention Gate	0.5593	0.9377	0.4446
U-Net + Halo Attention	0.4437	0.9343	0.332
U-Net + Triplet Attention	**0.807**	**0.9797**	**0.7226**
U-Net + Spatial Attention	0.5656	0.9626	0.4691

4) Hyperkvasir Segmentation Dataset

Each attention U-Net is also evaluated using the Hyperkvasir dataset, with detailed results presented in Table IV.

TABLE IV. RESULTS OF THE HYPERKVASIR SEGMENTATION DATASET

Model	*Dice*	*Accuracy*	*mIoU*
U-Net Baseline	0.6479	0.9134	0.5349
U-Net + SE	0.6997	0.9242	0.5879
U-Net + CBAM	0.7051	0.922	0.6002
U-Net + Coord Attention	**0.7338**	**0.9351**	**0.6501**
U-Net + Attention Gate	0.7297	0.9311	0.628
U-Net + Halo Attention	0.4766	0.8041	0.3433
U-Net + Triplet Attention	0.7048	0.9273	0.5973
U-Net + Spatial Attention	0.5304	0.8042	0.3938

For the **HyperKvasir** dataset, which contains 1,000 images of gastrointestinal conditions, the baseline U-Net achieved 0.6479 (Dice) and 0.5349 (mIoU). **U-Net + Coordinate Attention** improved these metrics, achieving the best scores, **0.7338 (Dice)** and **0.6501 (mIoU)**, due to its capabilities of extracting positional details, allowing the model to better confine lesions and disease-specific regions within the gastrointestinal tract. In addition, **U-Net + Attention Gate** also achieved competitive performance with **0.7297 (Dice)** and **0.628 (mIoU)**, due to its ability to recalibrate channels, to better extract crucial features across various gastrointestinal diseases.

5) ISIC 2017 Dataset

We showcase the efficiency of each attention U-Net for evaluating ISIC 2017 for skin lesion segmentation in Table V.

For the **ISIC** dataset, which consists of 2,594 dermoscopic images used for skin lesion segmentation, the baseline U-Net achieved 0.7133 (Dice) and 0.6216 (mIoU). **U-Net + Coordinate Attention** achieved the best scores **0.8073 (Dice)** and **0.7163 (mIoU**). Coordinate Attention helps by embedding positional information, which is particularly beneficial for ISIC's challenging lesions, where location and boundary details are crucial for accurate segmentation. This attention module enables the model to maintain spatial awareness while focusing on critical lesion regions, leading to better results compared to the baseline U-Net. Moreover, **U-Net + CBAM** obtained the second best results with **0.7906 (Dice)** and **0.7028 (mIoU**), due to CBAM's ability to recalibrate both spatial and channel feature maps, allowing the model to better capture crucial features in lesion segmentation where variability in lesion size and color presents significant challenges.

TABLE V. RESULTS OF THE ISIC 2017 DATASET

Model	***Dice***	***Accuracy***	***mIoU***
U-Net Baseline	0.7133	0.8953	0.6216
U-Net + SE	0.7525	0.9067	0.6598
U-Net + CBAM	0.7906	0.9159	0.7028
U-Net + Coord Attention	**0.8073**	**0.9231**	**0.7163**
U-Net + Attention Gate	0.7855	0.9118	0.6898
U-Net + Halo Attention	0.7124	0.8723	0.6015
U-Net + Triplet Attention	0.7856	0.9117	0.6895
U-Net + Spatial Attention	0.6003	0.8702	0.4956

6) Kvasir-instrument Segmentation Dataset

We demonstrate each network efficiency in instrument segmentation using the Kvasir-instrument dataset in Table VI.

For the **Kvasir-Instrument** dataset, which consists of 590 images of medical instruments in endoscopic procedures, the baseline U-Net achieved 0.8115 (Dice) and 0.7266 (mIoU). **U-Net + CBAM** acheved the best scores, **0.8570 (Dice)** and **0.7834 (mIoU)**, due to its combined channel and spatial attention strategies, which helped the model better localize the instruments and capture critical spatial features. This attention mechanism was especially useful in differentiating between the medical instruments and background tissue, leading to improved segmentation performance compared to all other attention U-Net methods. **U-Net + SE** also showed notable improvements with **0.8421 (Dice)** and **0.7646 (mIoU)**, benefiting from SE's ability to recalibrate the importance of feature channels, allowing the model to better distinguish between instruments and surrounding tissues. This is particularly important in endoscopic procedures where instruments may overlap or blend into tissue regions.

TABLE VI. RESULTS OF THE KVASIR-INSTRUMENT SEGMENTATION DATASET

Model	***Dice***	***Accuracy***	***mIoU***
U-Net Baseline	0.8115	0.9715	0.7266
U-Net + SE	0.8421	0.9777	0.7646
U-Net + CBAM	**0.8570**	**0.9773**	**0.7834**
U-Net + Coord Attention	0.7057	0.9536	0.6059
U-Net + Attention Gate	0.8401	0.9758	0.7674
U-Net + Halo Attention	0.6634	0.9613	0.5573
U-Net + Triplet Attention	0.8421	0.9773	0.7673
U-Net + Spatial Attention	0.7731	0.9689	0.6932

7) Retina Blood Vessel Dataset

The effectiveness of each network is further evidenced in Table VII for evaluating the Retina Blood Vessel dataset.

TABLE VII. RESULTS OF THE RETINA-BLOOD-VESSEL DATASET

Model	***Dice***	***Accuracy***	***mIoU***
U-Net Baseline	0.7377	0.9659	0.5899
U-Net + SE	0.7661	0.969	0.6267
U-Net + CBAM	**0.7956**	**0.9686**	**0.6626**
U-Net + Coord Attention	0.7572	0.9678	0.6159
U-Net + Attention Gate	0.752	0.968	0.6077
U-Net + Halo Attention	0.6544	0.9529	0.4884
U-Net + Triplet Attention	0.7741	0.9699	0.6349
U-Net + Spatial Attention	0.7719	0.9693	0.6332

For the **Retina Blood Vessel Segmentation** dataset, which consists of 100 retinal images used for blood vessel segmentation, the baseline U-Net achieved 0.7377 (Dice) and 0.5899 (mIoU). **U-Net + CBAM** obtained the best scores, **0.7956 (Dice)** and **0.6626 (mIoU**). The combination of channel and spatial attention in CBAM allowed the model to capture the refined details of the retinal vessels while enhancing both the critical channel features and spatial locations. **U-Net + Triplet Attention** illustrated the second best results of **0.7741** (**Dice**) and **0.6349 (mIoU**), owing to the Triplet Attention strategy's ability in extracting spatial, channel and temporal dependencies simultanously. This in turn helped the model better focus on the thin and elongated structures of the blood vessels, leading to better vessel delineation and detection.

D. Performance Comparison Across Datasets

We summarize model performance across all 7 datasets (i.e. Kvasir-SEG, Bowl, CVC-ClinicDB, Hyperkvasir, ISIC 2017, Kvasir-instrument, and Retina Blood Vessel datasets).

The mean Dice, mIoU and accuracy results across seven datasets for each model are provided in Table VIII. As indicated in Table VIII, the most performant top two networks are **U-Net + Triplet Attention and U-Net + SE**, owing to their great efficiency in extracting spatial, channel and temporal dependencies from a global perspective across all segmentation tasks. Figures 2-3 illustrate the example segmentation outcomes for these best attention networks for Kvasir-SEG, ISIC 2017 and Retina Blood Vessel datasets.

TABLE VIII. MEAN DICE, MIOU AND ACCURACY SCORES ACROSS 7 DATASETS

Model	*Dice*	*Accuracy*	*mIoU*
U-Net Baseline	0.6841	0.9334	0.5786
U-Net + SE	0.7687	0.9481	0.6643
U-Net + CBAM	0.7424	0.9432	0.6427
U-Net + Coord Attention	0.7139	0.9402	0.6115
U-Net + Attention Gate	0.7301	0.9369	0.6240
U-Net + Halo Attention	0.5920	0.9049	0.4718
U-Net + Triplet Attention	**0.7707**	**0.9467**	**0.6681**
U-Net + Spatial Attention	0.6722	0.9187	0.5606

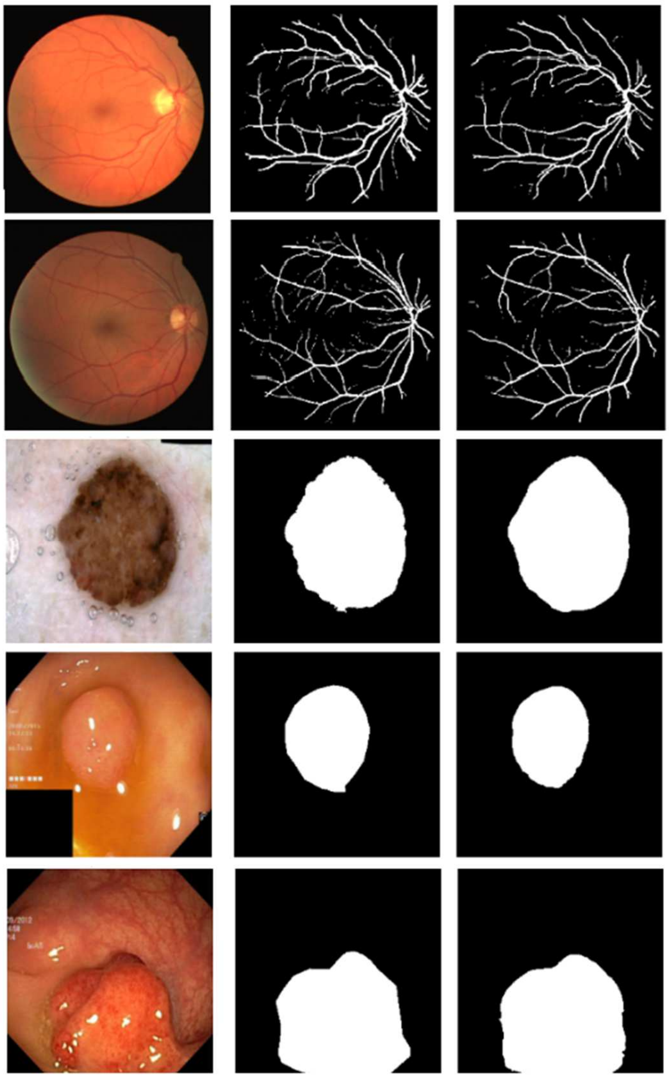

Fig. 2. Segmentation results for the proposed U-Net + Triplet Attention (from left to right, input images, GT and generated masks).

E. Comparison with Existing Studies

Several studies for polyp, skin lesion, nuclei and blood vessel segmentation with similar evaluation strategies are selected for performance comparison, as shown in Table IX.

This suggests that the introduction of attention mechanisms in our models allows for better focus on small and complex cellular, lesion and fine elongated structures, improving overall segmentation performance.

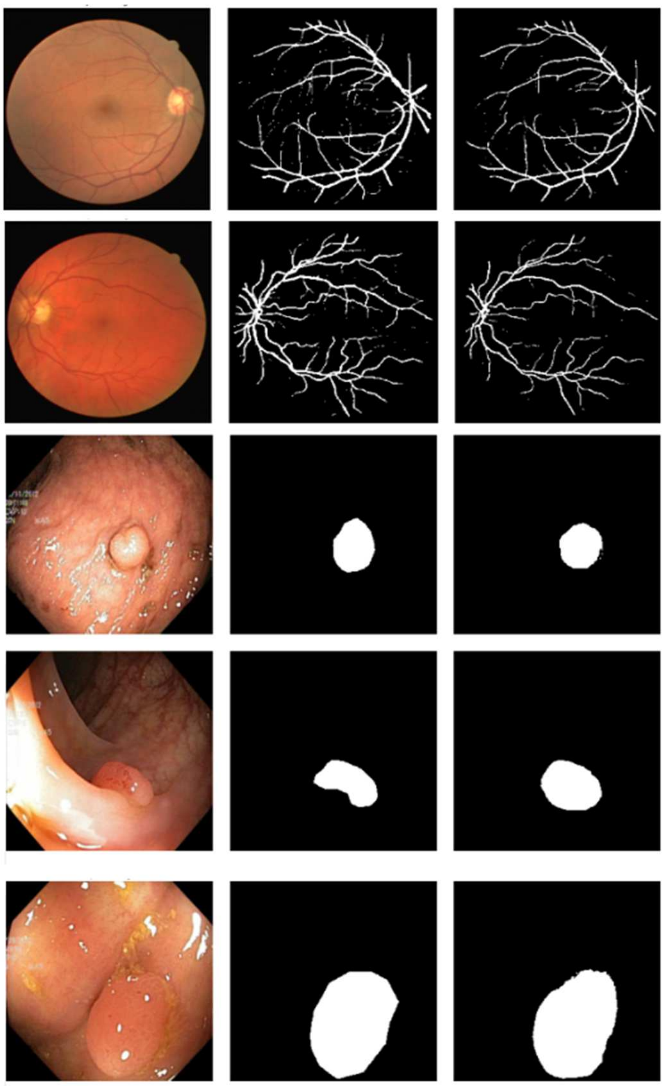

Fig. 3. Segmentation results for the proposed U-Net + SE Attention (from left to right, input images, GT and generated masks).

TABLE IX. COMPARISON WITH EXISTING STUDIES

Dataset	*Model*	*Dice*
Kvasir-SEG	Double U-Net [9]	0.8180
	U-Net + Attention Gate (ours)	**0.8959**
	U-Net + Coord Attention (ours)	**0.8850**
ISIC 2017	KM + deep network ensemble [10]	0.7315
	Deep ensemble networks [11]	0.7650
	U-Net + Coord Attention (ours)	**0.8073**
	U-Net + CBAM (ours)	**0.7906**
	U-Net + SE (ours)	**0.7525**
Data Science Bowl	Enhanced U-Net [12]	0.792
	U-Net + Attention Gate	**0.8325**
	U-Net + CBAM (ours)	**0.8118**
	U-Net + SE (ours)	**0.8035**
Retinal Vessel	CNN with attention [13]	0.741
	U-Net + CBAM (ours)	**0.7956**
	U-Net + SE (ours)	**0.7661**

V. Conclusion

In this research, we incorporated multiple attention modules, such as CBAM, Attention Gate, SE, Halo, Coordinate, Triplet and Spatial Attention mechanisms, into U-Net to enhance its ability to capture both local and global contextual information in medical images. The experimental results demonstrated significant improvements in segmentation accuracy, Dice, and mIoU scores across seven datasets, Kvasir-SEG, ISIC 2017, Data Science Bowl, CVC-ClinicDB, Kvasir Instrument, HyperKvasir and Retina Blood Vessel datasets.

The findings revealed that models utilizing attention modules, especially Triplet, SE, and CBAM, outperformed the baseline and other attention U-Net models in terms of Dice, mIoU and accuracy scores. Notably, attention mechanisms were particularly effective in handling complex structures with intricate boundaries, which is often crucial in medical imaging applications like polyp, skin lesion, medical instrument and retinal blood vessel segmentation. The ability of attention mechanisms to dynamically focus on the most relevant features contributed significantly to the superior performance of our networks. In future directions, we aim to incorporate MRI, CT scans, and ultrasound images, to allow attention-enhanced models to capture complementary details from different imaging contexts [14-18]. Another possible area of future work is the development of real-time and lightweight models suitable for deployment in clinical settings. Therefore, optimizing these models for efficiency while maintaining accuracy will be crucial for their practical use in healthcare [19-23]. Finally, clinical validation of these models is essential, along with a focus on improving interpretability of the proposed attention networks [24-27].

References

[1] Hu, J., Shen, L. and Sun, G., 2018. Squeeze-and-Excitation Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.7132-714.

[2] Woo, S., Park, J., Lee, J.Y. and Kweon, I.S., 2018. CBAM: Convolutional Block Attention Module. Proceedings of the European Conference on Computer Vision (ECCV), pp.3-19.

[3] Oktay, O., Schlemper, J., Le Folgoc, L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N., Kainz, B. and Glocker, B., 2018. Attention U-Net: Learning Where to Look for the Pancreas. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2018, pp.210-219.

[4] Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N. and Liang, J., 2019. UNet++: A nested U-Net architecture for medical image segmentation. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (pp. 3-11). Springer.

[5] Hou, Q., Zhou, D., Feng, J., Cheng, M.M. and Feng, J., 2021. Coordinate Attention for Efficient Mobile Network Design. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.13713-13722.

[6] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L. and Zhou, Y., 2021. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. arXiv preprint arXiv:2102.04306.

[7] Vaswani, A., Ramachandran, P., Srinivas, A., Parmar, N., Hechtman, B. and Shlens, J., 2021. Scaling local self-attention for parameter efficient visual backbones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12894-12904).

[8] Misra, D., Nalamada, T., Arasanipalai, A.U. and Hou, Q., 2021. Rotate to attend: Convolutional triplet attention module. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 3139-3148).

[9] Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., de Lange, T., Halvorsen, P. and Pogorelov, K., 2020. ResUNet++: An Advanced Architecture for Medical Image Segmentation. IEEE Access, 9, pp.122634-122646.

[10] Tan, T.Y., Zhang, L. and Lim, C.P., 2020. Adaptive melanoma diagnosis using evolving clustering, ensemble and deep neural networks. *Knowledge-Based Systems*, *187*, p.104807.

[11] Bi, L., Kim, J., Ahn, E., Feng, D. and Fulham, M., 2017. Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks. arXiv preprint arXiv:1703.04197.

[12] Gupta, V., Nguyen, T.V., Yeh, S.C., Gruev, V. and Theodore, N., 2018. Nuclei segmentation in histopathology images using deep fully convolutional networks. IEEE Transactions on Medical Imaging, 37(8), pp.1862-1871

[13] Ronneberger, O., Fischer, P. and Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 234-241). Springer, Cham.

[14] Hassanin, M., Anwar, S., Radwan, I., Khan, F.S. and Mian, A., 2024. Visual attention methods in deep learning: An in-depth survey. *Information Fusion*, *108*, p.102417.

[15] Chin Neoh, S., Srisukkham, W., Zhang, L., Todryk, S., Greystoke, B., Peng Lim, C., Alamgir Hossain, M. and Aslam, N., 2015. An intelligent decision support system for leukaemia diagnosis using microscopic blood images. *Scientific reports*, *5*(1), p.14938

[16] Zhang, L. and Lim, C.P., 2020. Intelligent optic disc segmentation using improved particle swarm optimization and evolving ensemble models. *Applied Soft Computing*, *92*, p.106328.

[17] Joy, C.P., Mistry, K., Pillai, G. and Zhang, L., 2022. A Hybrid-DE for Automatic Retinal Image-Based Blood Vessel Segmentation. In *Recent Advances in AI-enabled Automated Medical Diagnosis* (pp. 157-172). CRC Press.

[18] Lem, H. and Zhang, L., 2023, October. Mask R-CNN Transfer Learning Variants for Multi-Organ Medical Image Segmentation. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 1209-1216). IEEE.

[19] Tan, T.Y., Zhang, L., Lim, C.P., Fielding, B., Yu, Y. and Anderson, E., 2019. Evolving ensemble models for image segmentation using enhanced particle swarm optimization. *IEEE access*, *7*, pp.34004-34019.

[20] Slade, S., Zhang, L., Yu, Y. and Lim, C.P., 2022. An evolving ensemble model of multi-stream convolutional neural networks for human action recognition in still images. *Neural computing and applications*, *34*(11), pp.9205-9231.

[21] Slade, S., Zhang, L., Huang, H., Asadi, H., Lim, C.P., Yu, Y., Zhao, D., Lin, H. and Gao, R., 2023. Neural inference search for multiloss segmentation models. *IEEE Transactions on Neural Networks and Learning Systems*.

[22] Zhang, L., Slade, S., Lim, C.P., Asadi, H., Nahavandi, S., Huang, H. and Ruan, H., 2023. Semantic segmentation using Firefly Algorithm-based evolving ensemble deep neural networks. *Knowledge-Based Systems*, *277*, p.110828.

[23] Zhang, L., Lim, C.P. and Liu, C., 2023. Enhanced bare-bones particle swarm optimization based evolving deep neural networks. *Expert systems with applications*, *230*, p.120642.

[24] Cunha, L., Zhang, L., Sowan, B., Lim, C.P. and Kong, Y., 2024. Video deepfake detection using Particle Swarm Optimization improved deep neural networks. *Neural Computing and Applications*, *36*(15), pp.8417-8453.

[25] Slade, S., Zhang, L., Asadi, H., Lim, C.P., Yu, Y., Zhao, D., Panesar, A., Wu, P.F. and Gao, R., 2025. Cluster search optimisation of deep neural networks for audio emotion classification. *Knowledge-Based Systems*, *314*, p.113223.

[26] Zhang, L., Zhao, D., Lim, C.P., Asadi, H., Huang, H., Yu, Y. and Gao, R., 2024. Video deepfake classification using particle swarm optimization-based evolving ensemble models. *Knowledge-Based Systems*, *289*, p.111461.

[27] Raghavan, K., 2024. Attention guided grad-CAM: an improved explainable artificial intelligence model for infrared breast cancer detection. *Multimedia Tools and Applications*, *83*(19), pp.57551-5757.