

Attention mechanisms in UNet variants for medical/non-medical image segmentation: A comprehensive and state-of-the-art narrative review

Siddharth Gupta^a, Yogendra Chhetri^b, Amit Pandey^c, Akansha Singh^c,
Mrinalini Bhagawati^d, Suneet K. Gupta^{e,f}, Neha Gupta^{g,h}, Surbhi Verma^h,
Ekta Tiwariⁱ, Mustafa Al-Maini^j, Mostafa M. Fouda^k, Gavino Faa^l, Esma R. Isenovic^m,
Andrew Laineⁿ, Luca Saba^o, Jasjit S. Suri^{a,i,p,q,*}

^a Stroke Diagnostics and Monitoring Division, AtheroPoint™, Roseville, CA, 95661, USA

^b Department of Centre for Continuing Education, IIS, Bengaluru, Karnataka, 560012, India

^c Department of Computer Science, Bennett University, Gautam Buddha Nagar, UP, 201310, India

^d Department of Biomedical Engineering, Northeastern Hill University, Shillong, ML, 793022, India

^e Department of Computer Science and Engineering, Chandigarh University, Mohali, PB, 140413, India

^f VIZJA University, Warsaw, 00-854, Poland

^g IT Department, Bharati Vidyapeeth's College of Engineering, New Delhi, 110063, India

^h Department of CSE, KIET Group of Institutions, Ghaziabad, UP, 201206, India

ⁱ Department of Innovation, Global Biomedical Technologies, Inc., Roseville, CA, 95661, USA

^j Allergy, Clinical Immunology and Rheumatology Institute, Toronto, L4Z 4C4, Canada

^k Department of Electrical and Computer Engineering, Idaho State University, Pocatello, ID, 83209, USA

^l Department of Pathology, University of Cagliari, Cagliari, 09124, Italy

^m Department of Radiobiology and Molecular Genetics, National Institute of the Republic of Serbia, University of Belgrade, 11001, Serbia

ⁿ Departments of Biomedical and Radiology, Columbia University, New York, NY, 10027, USA

^o Department of Radiology, Azienda Ospedaliero Universitaria (A.O.U.), Cagliari, 09128, Italy

^p Department of ECE, Idaho State University, Pocatello, ID, 83209, USA

^q Department of Computer Science & Engineering, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Nagpur Campus, 440008, India

ARTICLE INFO

Keywords:

UNet
Convolution/non-convolution
transformers
Pruning
Attention
Segmentation

ABSTRACT

Background: This review examines enhancements to the U-Net's ability to represent complex spatial structures for segmentation through attention mechanisms. The study surveys applications across both medical and non-medical domains, focusing on cSAM, cCAM, CBAM, and their associated serial (sfSCAM) and parallel (pfSCAM) fusion architectures. Beyond accuracy, the analysis also considers backbone design, diversity of evaluation metrics, pruning strategies, and potential sources of bias.

Methodology: Following PRISMA guidelines, we systematically searched major academic databases using U-Net and attention-related keywords. The final corpus was constructed through stepwise exclusions (E1–E3) from an initial set of 1807 retrieved records. Spatial, channel, and hybrid attention modules were taxonomized; their placement within encoder, decoder, skip, and bottleneck blocks was identified; and application trends across domains such as liver, brain, skin, and retinal imaging were summarized.

Findings: Attention-augmented U-Net models consistently improved the segmentation of subtle, scale-dependent structures across modalities, including CT, MRI, ultrasound, dermoscopy, fundus, and satellite imaging. Spatial attention in skip connections and decoders reduced noise while preserving boundaries, whereas encoder-level channel attention improved semantic weighting; CBAM combined both effects. Parallel fusion (pfSCAM)

* Corresponding author. Stroke Monitoring and Diagnostic Division, AtheroPoint™, Roseville, CA, 95661, USA.

E-mail address: jasjit.suri@atheropoint.com (J.S. Suri).

¹ FIEEE Fellow, Institute of Electrical and Electronics Engineering.

² FAIMBE Fellow, American Institute of Medical and Biological Engineering.

³ FAIUM Fellow, American Institute of Ultrasound in Medicine.

⁴ FSVF Fellow, Society of Vascular Medicine.

⁵ FAPVS Fellow, Asia Pacific Vascular Society.

⁶ FAAAAI Fellow, Asia Pacific Association of Artificial Intelligence.

balanced accuracy and efficiency, while serial fusion (sfSCAM) enabled stable feature refinement. Attention placement strongly influenced global–local fusion and noise suppression. Pruning reduced computational cost with minimal performance loss, although evaluation remained focused on Dice and cross-entropy, with limited use of IoU, Hausdorff distance, and F1-score.

Conclusion: Spatial and channel attention, along with their serial and parallel fusion variants, function as structural design principles within the U-Net. Effective practice integrates careful module placement, pruning or quantization strategies, and explicit bias auditing.

1. Introduction

Over the past five decades, medical image segmentation has evolved through several methodological stages. From region-based approaches [1], to frontier-focused strategies like parametric-based, geometric-based level sets [2] [3], segmentation has evolved through multiple stages, culminating in the usage of artificial intelligence (AI) [4–6]. Sethian, who first presented the segmentation approach for the geometric-based level, and subsequently developed by Suri et al., for medical imaging. The level set-based approach relied on anchoring the final borders at edges with high gradients and on zero-level curve convergence via partial differential equations [7,8]. Level set-based geometric curves initially had issues due to manual initialization; these issues were later resolved through automation. However, different speed functions or regularization terms are needed to stop boundary bleeding in segmented organs [9]. Boundary bleeding resulted from factors such as noise, incomplete shape, and inadequate capture of target organs in pictures, which decreased performance metrics like the Dice similarity scores and the Jaccard index [10]. Efficiency-oriented architectures with integrated attention and full-scale skip connections, as ES-UNet presented by Park et al. [11], are encouraged by recent 3D segmentation research that has also brought attention to these problems when applying 2D UNet frameworks to volumetric data.

In medical image analysis, machine learning (ML) frameworks leverage annotated clinical datasets to automate diagnosis and segmentation tasks and apply that knowledge in classification or characterization frameworks [12,13]. Numerous studies on automated medical diagnosis are available [14–17], encompassing a range of image modalities as computed tomography [18], ultrasound [19], and magnetic resonance imaging (MRI) [20]. Despite the strength of these training models, they struggle to (a) extract training features in real time and (b) train optimization methods that are intended to reduce noise-induced bottlenecks when using gradient search to identify local minima [21]. As a result, machine learning (ML) dominated the field of artificial intelligence for a while until the deep learning (DL) paradigm's urgent need for automated feature extraction in segmentation emerged [5,17], as demonstrated by architectures such as UNet [22]. UNet's encoder-decoder design, first presented by Ronneberger et al. [22] in 2015, combines skip connections to fuse local and global contextual information. This improves segmentation accuracy by combining path features that expand and contract within a pruning framework that eliminates redundant weights during propagation.

Attention mechanisms were subsequently introduced into convolutional neural networks (CNNs) to enable them to focus on specific regions of the input data for image segmentation. An attention mechanism can be classified as either convolution-based or convolution-free. The first convolution-based attention mechanism is called cSAM. It alters the network's feature mappings by using a gating network to generate a spatial weight map. The introduction of attention gates in UNet was pioneered by Oktay et al. [23], were the first to introduce attention barriers in UNet in 2018. To address different sizes and capture local and global contexts, this invention integrated numerous cSAMs into the network's encoder and decoder paths. The second type, which focuses on channel attention, is the squeeze-and-excitation (SE) network, developed by Hu et al. [24]. Woo et al. introduced the convolutional block attention module (cCBAM), which is similar to sfSCAM [25]. A pfSCAM

mechanism was introduced by Hou et al. [26]. The Dual Attention UNet, introduced by Chen et al. [27] in 2021, is an illustration of attention mechanism fusion. This network can attend to the channel and spatial aspects of the feature maps (FM) because it incorporates both cCAM and cSAM into the encoder and decoder routes [28,29]. Vaswani et al. [30] introduced the Transformer, a convolution-free architecture built entirely on self-attention mechanisms [31].

Given the architectural dominance of UNet in medical image segmentation and the extensive integration of attention mechanisms within its encoder–decoder framework, this review deliberately focuses on attention-enhanced UNet variants. This targeted scope enables structured, placement-centric analysis and extraction of architectural design principles, rather than presenting a broad but less cohesive survey across heterogeneous backbone architectures. This review is structured based on the functional components of the UNet architecture, namely the encoder, skip connections, decoder, and bottleneck. Within each component, attention mechanisms are further categorized according to their type, including spatial, channel, and hybrid fusion strategies. In addition, application-oriented classification is provided in [Appendix C](#) to relate architectural design choices with specific medical segmentation tasks.

The literature review, which focuses on the development of attention mechanisms in UNet-based architectures for medical picture segmentation, is presented in Section 2. The statistical distribution and search approach are shown in Section 3. Additionally, the classification of attention mechanisms is covered in part 4, while sections 5, 6, 7, and 8 detail how different attention mechanisms affect the UNet's arms (encoder, decoder, skip, and bottleneck). The critical discussion in Section 9 covers the planned study's merits, flaws, and benchmarking. In Section 10, we finally wrap up the suggested investigation. Additionally, [Appendix A](#) discusses transformers, which are convolutional free attention models. [Appendix B](#) discusses how different attention mechanisms affect the multiarm UNet, while [Appendix C](#) discusses a single issue are numerous solutions.

2. Search methodology and statistical analysis

Understanding distinct attention mechanisms and how they affect specific UNet components in both medical and non-medical contexts requires a statistical analysis of the literature. To comprehend how different attention mechanisms affect the various parts of UNet, the following parameters are required. We have chosen a practical, robust implementation and sustainability methodology (PRISMA) for choosing research on attention mechanisms in UNet and evaluating bias in order to overcome the following problems. In particular, the statistical analysis is presented in Section 2.2, and the research selection procedures are covered in Section 2.1. The first is the use of feature extraction techniques; the second is the variety and consistency of the performance metrics used in the selected studies; the third is the use of pruning models for storage optimization; and the fourth is the bias present in AI-based solutions.

2.1. Study selection using PRISMA model

The studies included in this review were chosen and found using the PRISMA approach. The search terms included "Attention UNet for non-

medical studies," "Attention UNet for medical studies," "UNet and Attention for segmentation," "Attention UNet-based segmentation," "Self Attention and UNet," "UNet and Transformer," "UNet and self-attention based segmentation," "Channel attention and UNet-based segmentation," "Spatial Attention UNet-based segmentation," "cCBAM and UNet-based segmentation," "AI-Bias in Attention UNet," "Pruning methods for attention and UNet," and "Storage size of Attention and UNet." The search was conducted across several databases, including ScienceDirect, IEEE Xplore, Scopus, PubMed, Web of Science, and Google Scholar. The PRISMA-compliant flow chart for the carefully chosen studies is shown in Fig. 1. After completing the database search, a total of 1807 records were identified (PubMed: 112; Google Scholar: 1445; ScienceDirect: 235; IEEE Xplore: 15). After duplicate removal, 1612 unique records remained. During the title and abstract screening phase, 1410 non-relevant articles were excluded (E1), resulting in 202 records progressing to detailed evaluation. Of these, 160 articles met the criteria for full-text assessment. During the eligibility stage, 92 records were excluded after detailed screening (E2), and an additional 11 studies were excluded due to insufficient methodological or experimental detail (E3). Ultimately, 149 studies were included in the qualitative synthesis. The exclusion categories (E1, E2, E3) are explicitly illustrated in Fig. 1 to ensure transparent traceability of the screening process.

2.2. Statistical distribution

The statistical distribution was divided into four sections. First, as illustrated in Fig. 2(a), we separated the attention types into four categories: spatial, channel, sfSCAM, and pfSCAM. In the second section, as illustrated in Fig. 2 (b). As shown in Fig. 2(c), we separated the papers into convolution-based and non-convolution-based attention applications in the third section. Lastly, as illustrated in Fig. 2 (d), we separated the quantity of papers according to the application, such as skin, liver, and cell nuclei. Overall, this statistical breakdown provides a comprehensive view of how attention mechanisms are distributed across different architectural designs and application domains. The analysis highlights prevailing research trends revealing underexplored areas that may benefit from further investigation. Such insights can help guide

future developments in attention-based medical image analysis.

3. Literature review

Prior to 2014, sequence models that processed data step-by-step, such as recurrent neural networks (RNNs) [32] and long short-term memory LSTM [33] were unable to grasp long-range dependencies. In order to overcome this constraint, Bahdanau et al. [34] proposed the soft attention mechanism in 2014, which gave each word a weight based on its significance. This reduced computational complexity and enabled more intelligent alignment between the source and destination texts.

Between 2015 and 2018, attention gates were frequently employed in tasks such as image captioning [35], machine translation [34], and classification [36]. Attention mechanisms are generally classified into three primary categories: hard attention [37], soft attention [34], and self-attention, also known as transformer attention [38]. Hard attention relies on pruning techniques, whereas soft attention is founded on the principle of backpropagation. Soft attention mechanisms are further classified into spatial attention [39], channel attention [24], and the fusion of channel and spatial attention [25]. In 2017, Vaswani et al. [38] proposed a paper entitled "Attention is all you need." This paper entirely abandoned the notion of RNNs, instead adopting a self-attention feed-forward architecture and introducing the concept of transformers. Self-attention enabled the model to concurrently compare all positions within a sequence, thereby facilitating the efficient processing of long-range dependencies. It also facilitated parallel processing, thereby significantly increasing training speed. Transformers formed the foundational model for GPT [40], BERT [41], and vision transformer [42] after 2018. The transformer architecture incorporates a multi-head self-attention mechanism, enabling the model to analyze input from diverse viewpoints. Cross attention [42], sparse attention [43], and retrieval-based attention, like GPT4 [43], are examples of more sophisticated attentional evolutions in this age.

Ronneberger et al. presented UNet, a deep learning model with a U-shaped architecture, in 2015 [22]. The preliminary layers of the UNet architecture contain rudimentary feature representations. Propagating these features through skip connections from encoder to decoder, across

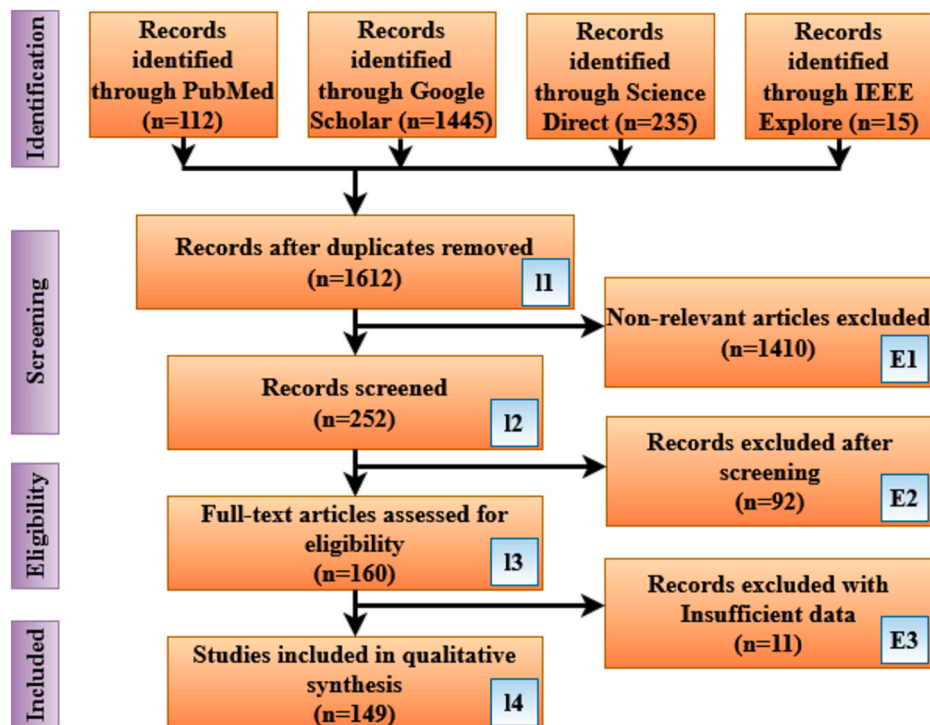


Fig. 1. PRISMA model for study selection.

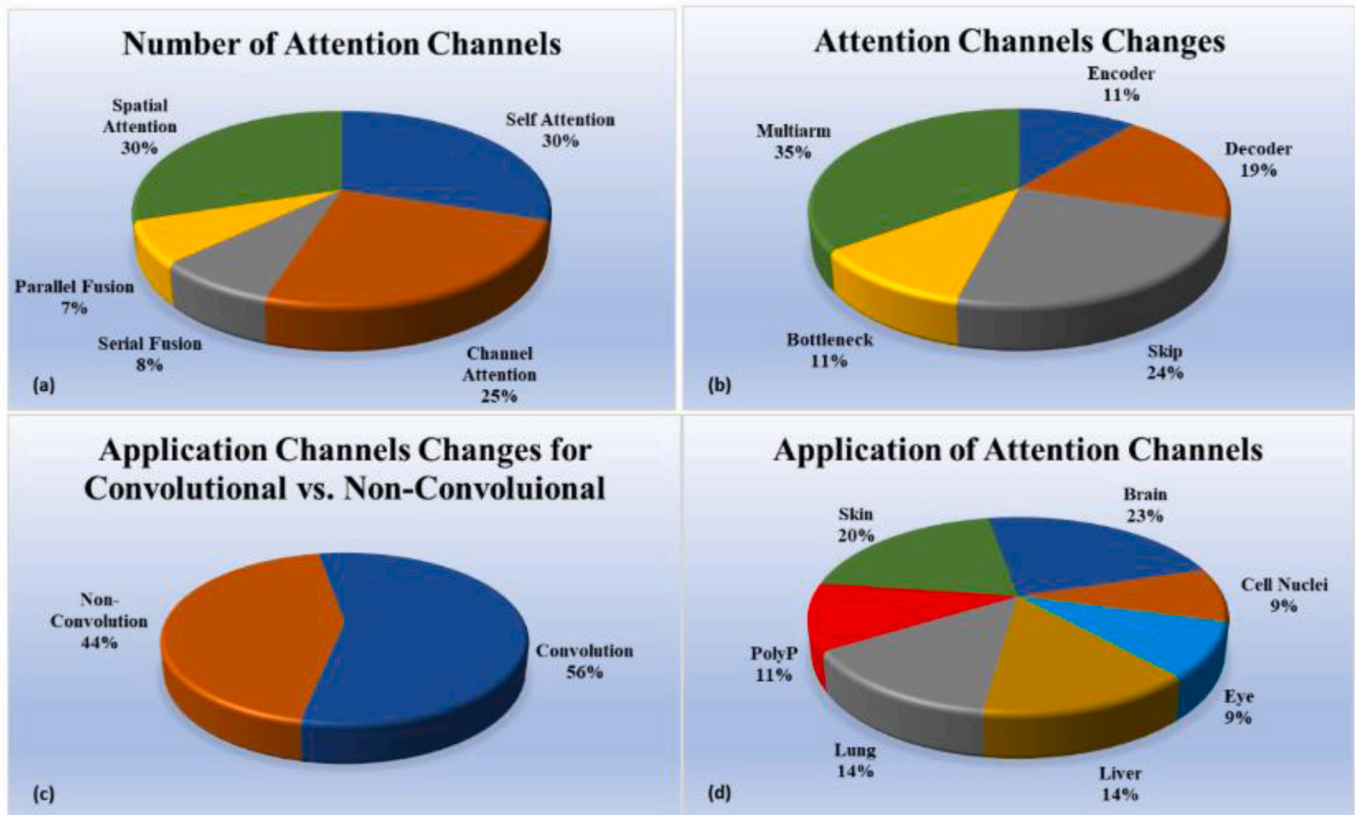


Fig. 2. Statistical distribution (a) Number of Attention Channels; (b) Changes in attention channels were observed across various model segments, including the Encoder, Decoder, Skip Connection, Bottleneck, and Multi-arm components. (c) Attention channel changes for Convolution vs. Non-Convolution applications; and (d) Attention channel changes by applications.

encoders, or between decoders proved inefficient and resource-intensive. The encoder, decoder, skip connections, bottleneck, and other UNet arms can employ attention methods to solve this issue. Recent research has focused on refining this integration. For instance, Pan et al. [44] presented a multi-scale convolutional attention UNet that improves boundary precision by adaptively fusing local and global contextual inputs. By incorporating multi-head and multi-scale cross-axis attention modules that improve feature correlation across encoder-decoder levels, Yan et al. [45] expanded on this concept with MDEUNet. While Al-Ebrahim [46] investigated spike-based attention to improve temporal and spatial concentration in biomedical segmentation, lightweight adaptations like AttE-UNet [47] show that attention augmentation can improve segmentation accuracy even in limited situations. These advancements demonstrate how attention processes continue to evolve, enabling context-aware, more effective UNet variants for medical image segmentation.

Despite the architectural strengths of the original UNet, several computational bottlenecks emerged during its widespread adoption in medical image segmentation. First, direct skip connections propagated both relevant and redundant low-level features, potentially introducing noise and reducing boundary precision. Second, convolutional operations rely on local receptive fields, limiting efficient modeling of long-range spatial dependencies and global anatomical context. Third, deep convolutional networks often suffer from channel redundancy, where semantically weak or overlapping feature maps contribute equally during decoding. Finally, scaling UNet architectures for volumetric and high-resolution medical data increases memory consumption and computational burden. These limitations motivated the integration of attention mechanisms as adaptive gating strategies to selectively enhance informative features, suppress irrelevant activations, and improve global-local feature interaction without excessive architectural

expansion.

4. Classification of attention models

Advancements in AI, particularly in the realm of deep learning, have propelled significant innovations in medical imaging, with attention models emerging as a pivotal area of research. These models, which enhance feature extraction and improve interpretability by focusing on relevant parts of input data, have become increasingly important in addressing complex challenges in medical diagnostics. This chapter delves into the sophisticated architectures and mechanisms of attention models, focusing on their integration into UNet for biomedical image segmentation. We begin by exploring the basic UNet architecture, a fundamental encoder-decoder structure optimized for detailed medical image analysis. We subsequently discuss the role of attention mechanisms that refine this model for enhanced performance. These attention mechanisms, both convolution-based and convolution-free, represent a critical evolution in the model's capability to discern and highlight salient features in medical images, thereby pushing the boundaries of what automated systems can perceive and interpret.

4.1. Basic UNet architecture and its components

UNet is a CNN architecture specifically designed for image segmentation tasks. UNet originated from a seminal publication, "UNet: Convolutional Networks for Biomedical Image Segmentation" by Ronneberger et al. [24], which was published in 2015. The encoder-decoder architecture of UNet is shown in Fig. 3. The encoder segment extracts the input image's high-level features, which the decoder segment uses to create a segmentation mask. Several convolution layers are usually used in the encoder section of the network, which

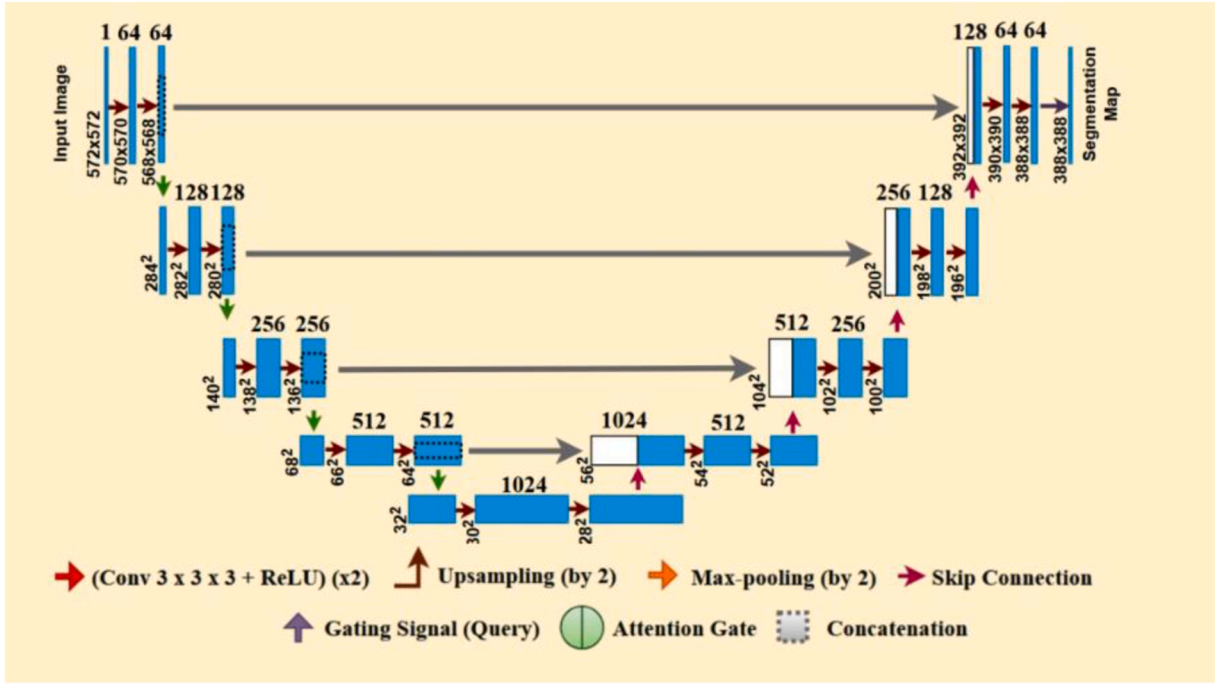


Fig. 3. Basic UNet architecture. Conv: Convolution; ReLU: Rectified Linear Unit.

continuously add new channels while progressively reducing the spatial resolution of the input image. Conversely, in the decoder segment of the network, transposed convolutional layers are utilized to up-sample the FMs, restoring them to the original input size while concurrently reducing the number of channels. UNet's pivotal innovation lies in its incorporation of skip connections linking the encoder and decoder layers. The encoder's high-resolution features are accessed through the decoder's connection, enhancing the quality of the segmentation mask. Consequently, UNet has emerged as a favored architecture for various image segmentation tasks, especially in the realm of medical image analysis.

4.2. Attention mechanism

In this work, the attention mechanism is divided into two main categories: convolution-based and convolution-free. The convolution-based method highlights relevant activations during training, thereby reducing computational resources spent on irrelevant activations and enhancing network generalization. Conversely, the convolution-free approach achieves similar objectives without relying on convolution operations. Inspired by the work of Oktay et al. [23], Hu et al. [24], and Woo et al. [25], we divide the convolution-based attention mechanism into four parts. This attention will be discussed in the following subsections, and the impact of attention mechanisms on different arms of UNet will be discussed in later sections.

4.3. Convolution-based spatial attention mechanism

The cSAM generates a spatial attention map by considering the inter-spatial relationships among features. cSAM highlights the 'where' of the information. To calculate the spatial attention using cSAM, we rely on five key components: The input feature, average-pooling (AP), max-pooling (MP) operations along the channel axis, and concatenation are combined to provide an efficient feature descriptor. Emphasizing informative regions has been successfully achieved using a procedure along the channel axis. The third step involves applying a convolution layer to the concatenated FM to generate a cSAM. Following this, a sigmoid activation function is applied to normalize the spatial attention

map. Let $F \in \mathbb{R}^{C \times H \times W}$ denote the input feature map extracted from the convolutional encoder, where C represents the number of channels and H, W denote spatial dimensions. The average-pooling (F_{AP}^s) and max-pooling (F_{MP}^s) operations are performed along the channel axis to generate spatial descriptors that summarize inter-spatial information. These descriptors are concatenated and passed through a learnable 7×7 convolution layer to model spatial relationships. A sigmoid activation function $\Theta(\cdot)$ is then applied to normalize the resulting spatial attention map between 0 and 1.

$$cSAM(F) = \Theta(f^{7 \times 7}([AP(F); MP(F)])) = \Theta(f^{7 \times 7}([F_{AP}^s; F_{MP}^s])) \quad (1)$$

4.4. Convolution-based channel attention mechanism

The convolution-based channel attention mechanism (cCAM) produces a channel attention map by leveraging inter-channel feature relationships. Each channel of the FM is considered a feature detector, allowing cCAM to focus on "WHAT" is meaningful in the given input image. cCAM is fundamentally based on five pillars. The first is the input image. The second involves pooling to aggregate spatial information, using either AP, MP, or both. The third pillar is a shared network, specifically a shared multilayer perceptron (cMLP) with 1×1 convolutions and a single hidden layer aimed at reducing the number of parameters. The fourth pillar uses a sigmoid activation function to normalize the channel attention weights, as shown in Fig. 4. In the channel attention mechanism, the input feature map $F \in \mathbb{R}^{C \times H \times W}$ is first aggregated using average-pooling (F_{AP}^c) and max-pooling (F_{MP}^c) operations to produce channel descriptors. These descriptors are passed through a shared multilayer perceptron (cMLP) consisting of two learnable weight matrices W_{T0} and W_{T1} , where a reduction ratio r is used to reduce parameter complexity. The first transformation reduces the channel dimension, followed by a ReLU activation introducing non-linearity, and the second transformation restores the channel dimension. A sigmoid activation function $\Theta(\cdot)$ normalizes the resulting channel attention weights between 0 and 1. The mathematical explanation of cCAM is as follows:

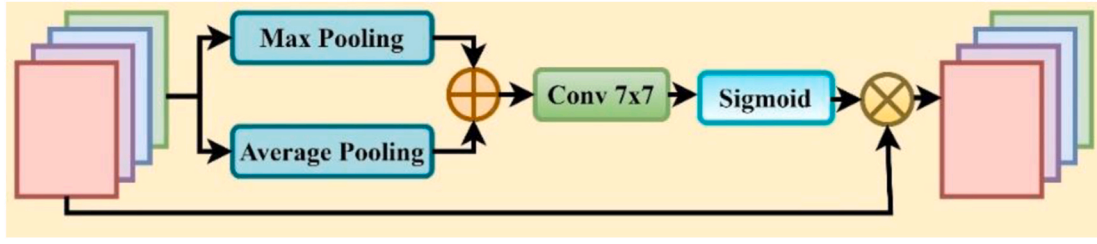


Fig. 4. Architecture of cSAM.

$$\begin{aligned} \text{cCAM}(\mathbf{F}) &= \Theta(\text{cMLP}(\text{AP}(\mathbf{F})) + \text{cMLP}(\text{MP}(\mathbf{F}))) \\ &= \Theta(\text{WT}_1(\text{ReLU}(\text{WT}_0(\mathbf{F}_{\text{AP}}))) + \text{WT}_1(\text{ReLU}(\text{WT}_0(\mathbf{F}_{\text{MP}})))) \end{aligned} \quad (2)$$

where Θ represents the sigmoid activation function, $\text{WT}_0 \in \mathbb{R}^{C \times r}$ and $\text{WT}_1 \in \mathbb{R}^{C \times r}$. Note that the cMLP weights WT_0 and WT_1 are shared for inputs and the Relu activation function, followed by WT_0 .

4.5. Sequential integration of spatial attention and channel attention

Consider $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$, an FM serving as an input feature representation where C indicates the channel count, H specifies the height, and W denotes the width. sfSCAM refers to the fusion of one-dimensional cCAM, $\text{cCAM} \in \mathbb{R}^{C \times 1 \times 1}$, 5 shown in Fig. 5, and two-dimensional cSAM, $\text{cSAM} \in \mathbb{R}^{1 \times H \times W}$ cSAM as described in Fig. 4. The final architecture is shown in Fig. 6, and the final FM is presented in Equations (3) and (4) below.

$$\mathbf{F}' = \text{cCAM}(\mathbf{F}) \odot \mathbf{F} \quad (3)$$

$$\mathbf{F}'' = \text{cSAM}(\mathbf{F}') \quad (4)$$

where $\mathbf{F}''$ is the final FM and $\mathbf{F}'$ is the FM after cCAM, $\odot$ represents element-wise multiplication. cSAM($\mathbf{F}'$) generates a spatial attention map from $\mathbf{F}'$, which is applied to $\mathbf{F}'$ via element-wise multiplication to obtain the final feature map $\mathbf{F}''$.

4.6. Parallel/concurrent integration of spatial attention and channel attention

Let $\mathbf{M}$, $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$ serve as an input feature representation, C indicates the channel count, H specifies the height, and W denotes the width. pfSCAM refers to one-dimensional cCAM, $\text{cCAM} \in \mathbb{R}^{C \times 1 \times 1}$, and two-dimensional cSAM, $\text{cSAM} \in \mathbb{R}^{1 \times H \times W}$, is illustrated in Fig. 7. The resulting output feature representation is depicted in Equation (5).

$$\mathbf{F}' = \text{cCAM}(\mathbf{F}) \odot \mathbf{F}$$

$$\mathbf{F}'' = \text{cSAM}(\mathbf{F}') \odot \mathbf{F}'$$

$$\mathbf{F}''' = \mathbf{F}' \odot \mathbf{F}'' \quad (5)$$

where $\mathbf{F}'$ is FM after cCAM, $\mathbf{F}''$ is FM after cSAM, and $\mathbf{F}'''$ is final FM, $\odot$ represents the concatenation of FMs.

5. Impact of the attention mechanism in the encoder of UNet

This section critically evaluates the influence of various attention mechanisms—cSAM, cCAM, sfSCAM, and pfSCAM—within the encoder of the UNet (eUNet) framework. It offers a comprehensive analysis, culminating in a discussion of specific applications where these mechanisms have been implemented, underscoring their impact on the encoder's functionality (refer to Table 1). Additionally, the examination extends to the image modalities and loss functions employed in these applications, providing a thorough understanding of how attention mechanisms optimize the encoder's performance in complex image segmentation tasks. This analysis not only showcases the practical applications but also elucidates the underlying modifications to the encoder structure that enhanced segmentation accuracy and efficiency.

5.1. Impact of convolution-based channel attention module on the encoder of UNet

The UNet network shown in Fig. 8 is mainly based on two components: an encoder-decoder for semantic segmentation and a shape stream that focuses on the boundary level. Every level of the encoder consists of BN [45], ReLU [46], and two 3x3 convolutions with a stride of 2; after that, residual blocks were used [54]. Training a deep neural network (DNN) is complex because the input parameters change earlier layers are adjusted. This process necessitates a lower learning rate and meticulous parameter initialization, slowing down the training. The solution to the problem is normalizing layer inputs. During training, DNN models encounter the issue of vanishing or exploding gradients. [55–57], which is solved by residual blocks [54].

The residual-block outputs are forwarded to the corresponding shape stream, where each input is processed with a 1×1 convolution to reduce channel dimensionality and introduce non-linearity before entering the attention block. Outputs of the first two attention layers are fed back into their respective residual blocks. For the final stage, the outputs of the

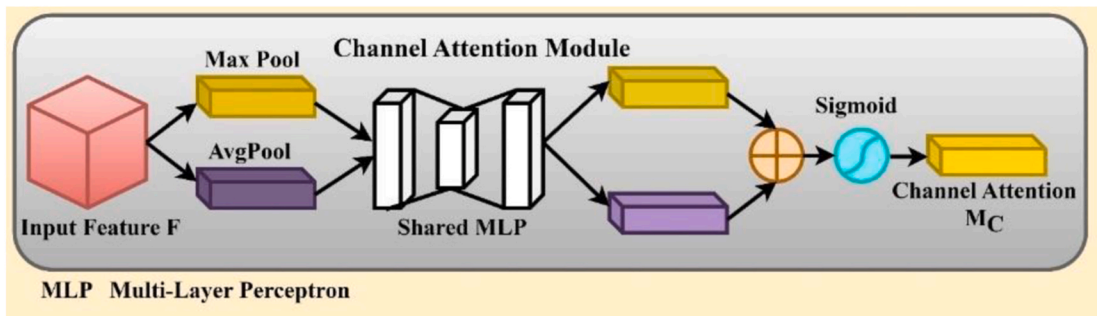


Fig. 5. Architecture of cCAM.

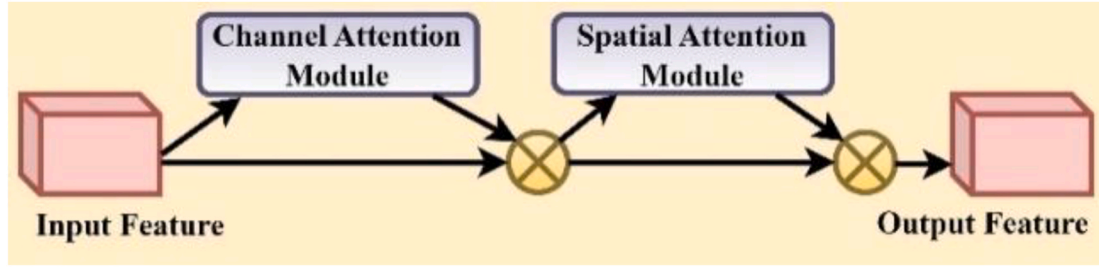


Fig. 6. Architecture of sfSCAM.

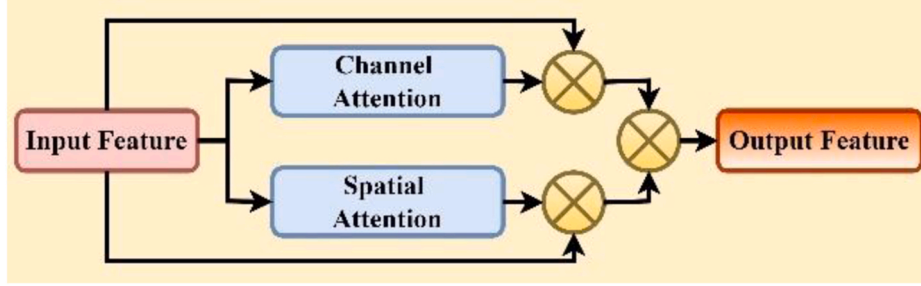


Fig. 7. Architecture of pfSCAM.

Table 1

Segmentation of medical images with the help of an attention mechanism on UNet. FOV: Field of View, LF: Loss Function, Attn.: Attention. CT: Computed Tomography, MRI: Magnetic Resonance Imaging, US: Ultrasound, pfSCAM: parallel fusion of Spatial and Channel Attention Module, cSAM: convolution-based Spatial Attention Module, cCAM: Convolution Based Channel Attention Module, sfSCAM: serial fusion of Spatial and Channel Attention Module; BCE: Binary Cross Entropy; RMSE: Root Mean Squared Error.

SN	Author/citation	Image Modality	FOV (Application)	Attn. Type	LF Type
1	Hou et al. [26]	CT	Aortic/Lung/Liver	pfSCAM	BCE
2	Hatamizadeh et al. [48]	MRI	Brain Tumor	cSAM	BCE/Dice
3	Qiao et al. [49]	CT	TCIA	cCAM	RMSE/Perceptual
4	Li et al. [50]	Dermoscopy	Skin lesion	sfSCAM	BCE/Dice
5	Yin et al. [51]	CT/Satellite/Camera	Cloud/Snow detection	sfSCAM	BCE
6	Singh et al. [52]	US	Antral follicle count	cCAM	BCE
7	Wang et al. [53]	MRI	Eye	pfSCAM	BCE

third attention layer and the last encoder layer are concatenated and passed to the ASPP module [58]. ASPP resamples features at multiple dilation rates to enlarge the effective field of view and capture multi-scale context. Its output then enters the expansive path, which applies 1×1 convolutions and bilinear upsampling at each layer to restore the original image resolution.

Each attention layer receives two inputs: the previous decoder feature x_l and the corresponding encoder feature x_g . These are concatenated, processed with a 1×1 convolution, and passed through a sigmoid activation to generate the attention map [59]. The map is then concatenated with x_l . This mechanism suppresses irrelevant regions and highlights informative structures. The attention map AM is computed as in Equations (6) and (7):

$$AM = \Theta(\text{conv}_{1 \times 1}(\odot x_g)) \quad (6)$$

$$O = AM \odot x_l \quad (7)$$

where $\odot$ represents the concatenation operation, and Θ represents the sigmoid activation function.

5.2. Impact of convolution-based channel attention module on the encoder of UNet

The architecture of UNet is shown in Fig. 9 and suggested by Jiang et al. [60], consists of eight residual blocks—four for down-sampling, four for squeeze-and-excitation (SE), and one for ASPP. Each residual unit applies a convolution followed by batch normalization and Leaky ReLU, providing mild regularization and reducing sensitivity to initialization. Down-sampling blocks use a 3×3 convolution with BN and Leaky ReLU, followed by 2×2 max-pooling. Up-sampling is performed with stacked 2×2 transposed convolutions and 3×3 convolutions. After each encoder convolution, a lightweight SE module enhances feature extraction by reweighting channels, allowing the network to focus on task-relevant structures. This improves fine-detail segmentation by reinforcing informative features at each residual stage. The ASPP block captures multi-scale context and expands the effective receptive field, reducing fragmented segmentation and improving robustness to the varied appearance of Greenland lakes. Skip connections transfer encoder features directly to matching decoder blocks, restoring spatial detail and supporting multi-scale prediction and deep supervision.

As shown in Fig. 10, the SE module operates on feature maps of size $H \times W \times C$. The squeeze step applies global average pooling, compressing each channel into a $1 \times 1 \times C$ descriptor $Z \in \mathbb{R}^C$ by aggregating spatial information across $H \times W$, as expressed in the equation below.

$$Z_c = F_{sq}(U_c) = \frac{1}{W \times H} \sum_{i=1}^W \sum_{j=1}^H F_c(i, j) \quad (9)$$

where $i \in W, j \in H$, and F_{sq} is FMs from the squeeze module. To prevent the model from becoming overly complex and to enhance generalization, two fully connected layers are introduced around the non-linearity in the excitation module. These layers act as a bottleneck to benchmark the pairwise gate mechanism. The mathematical equation describing this process is as follows:

$$S = F_{ex}(z, W) = \Theta(W_2 \delta(W_1 z)) \quad (10)$$

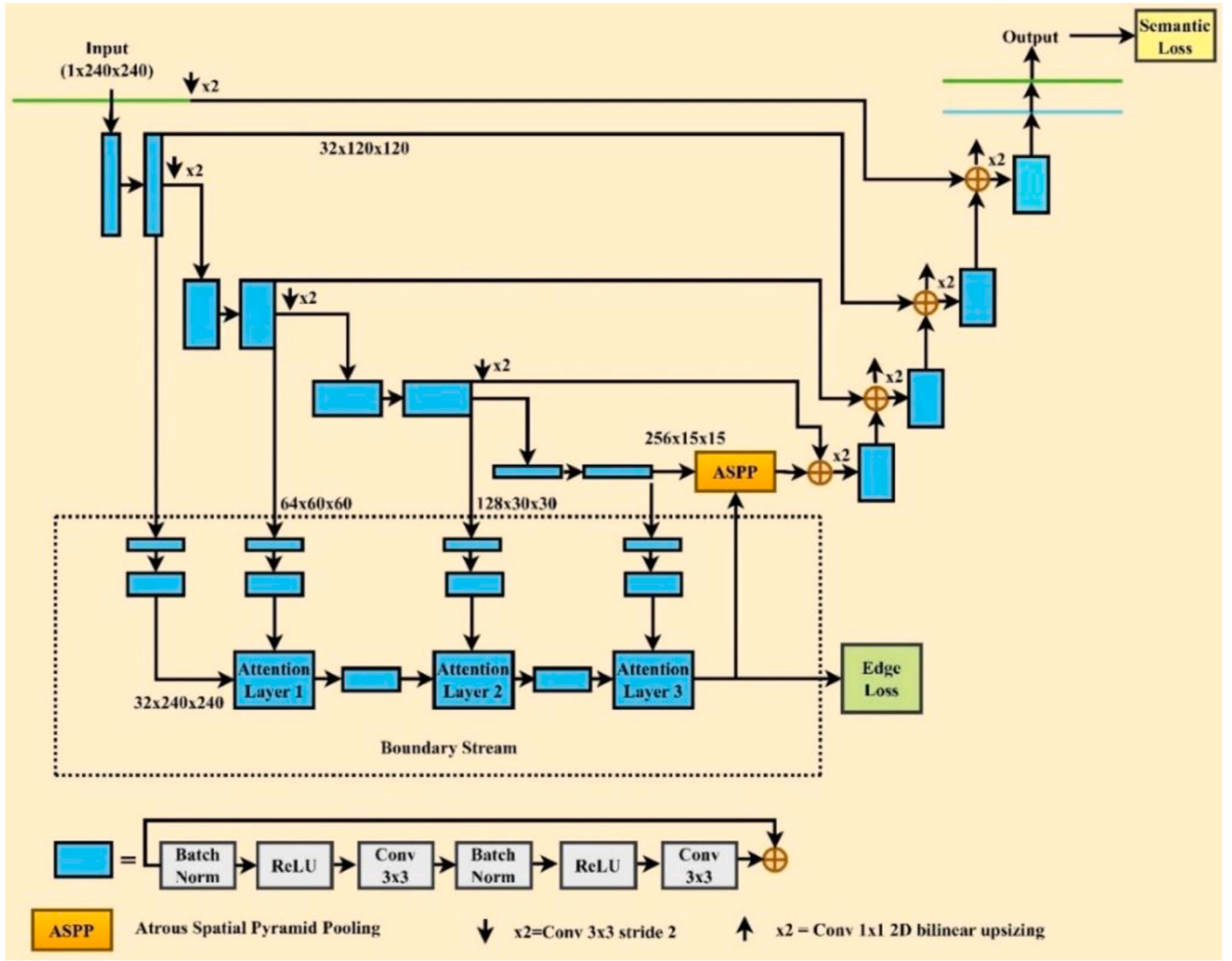


Fig. 8. Impact of cSAM on eUNet and atrous spatial pyramid pooling stages.

where δ represents the ReLU function and $W_1 \in R^{r \times r}$ and $W_2 \in R^{c \times r}$. and F_{ex} is FMs from the excitation module.

Squeeze-and-Excitation (SE) modules introduce additional parameters through channel reduction and expansion operations, with parameter growth proportional to $2C^2/r$, where C is the number of channels and r is the reduction ratio. When applied across multiple encoder layers, this overhead increases linearly with depth but remains relatively modest compared to convolutional layers, particularly for commonly used reduction ratios (e.g., $r = 8$ or 16). From a memory perspective, SE modules introduce negligible activation overhead since they operate on channel descriptors rather than full-resolution feature maps. Consequently, SE blocks are generally considered lightweight, though their cumulative impact may become noticeable in deep encoder configurations or high-resolution 3D segmentation settings.

5.3. Impact of serial fusion of spatial and channel attention module on the encoder of UNet

Yin et al. [51] replaced the UNet3+ feature extractor with a ResNet50-based backbone, as illustrated in Fig. 11. The original UNet3+ consists of five encoder blocks, each composed of three convolutional layers and a max-pooling layer. In the improved architecture, five ResNet50 encoder stages are used instead, incorporating convolutions

with multiple kernel sizes, regularization layers, ReLU activations, and residual modules, producing five effective feature maps (FM1–FM5). Each decoder layer applies a convolution followed by an activation function. Multi-scale fused feature maps (FFM1–FFM4) are then generated via skip-connection fusion, allowing the network to capture fine details while preserving broader contextual information.

Remote sensing images often contain large, irrelevant regions that consume unnecessary computation during feature extraction. To focus the model on cloud and snow regions, a convolutional block attention module (CBAM) is integrated into the ResNet50 backbone—also shown in Fig. 11. The channel attention module (cCAM) reweights feature channels to emphasize cloud- and snow-relevant information, while the spatial attention module (cSAM) highlights their spatial locations. Together, these mechanisms significantly enhance the network's ability to extract discriminative features from cloud and snow regions in remote sensing imagery.

5.4. Impact of parallel fusion of spatial and channel attention module on the encoder of UNet

Hou et al. [26] proposed an AF-NET framework illustrated in Fig. 12 which is structured around an encoder-decoder architecture enhanced with attention and feature fusion mechanisms. The framework comprises two main components: an encoder featuring DA blocks and a

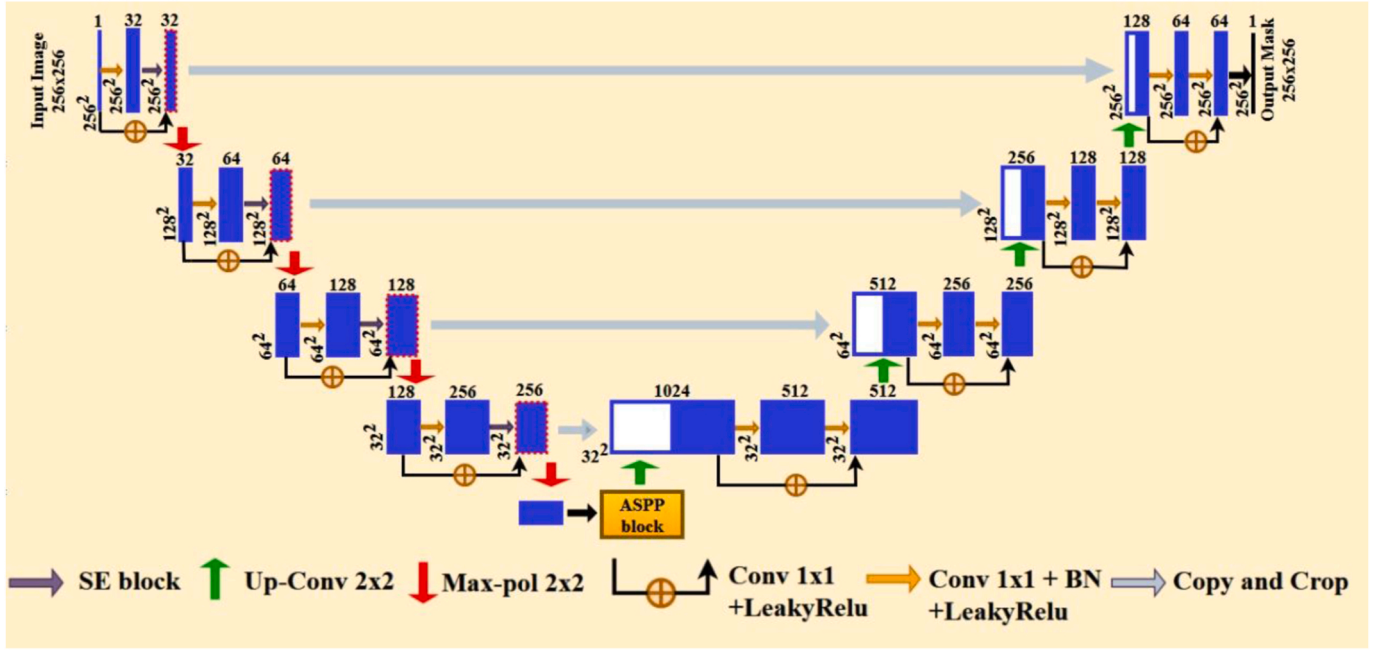


Fig. 9. Impact of cCAM on eUNet and ASPP.

SE: Squeeze-and-Excitation, ASPP: Atrous spatial pyramid pooling.

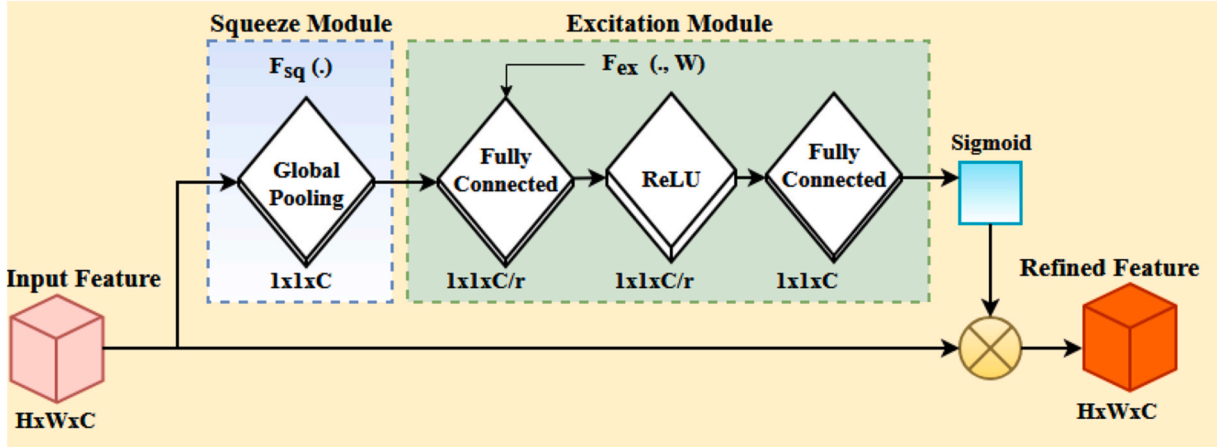


Fig. 10. Architecture SE and rectified linear unit (ReLU) are used as linear activation functions.

decoder incorporating MFF and GFF blocks. The DA (Dual Attention) blocks, depicted in Fig. 13, filter and weight the preprocessed image, generating multi-scale features essential for subsequent feature fusion. The architecture of the pfSCAM, illustrated in Fig. 14, further enhances the attention mechanism by selectively focusing on spatial and channel-wise information, thereby refining feature representation. These blocks enhance the network's ability to focus on relevant information and suppress irrelevant noise, thereby improving the quality of feature extraction. Following the DA blocks, the MFF (Multi-level Feature Fusion) extracts deeper global information from the enhanced features. This step aims to capture broader contextual information, which is crucial for accurate pixel localization and segmentation. Finally, the framework integrates the multi-scale features from the encoder with the intermediate results from the decoder. This integration process enables precise pixel-level predictions by leveraging local details and global context extracted throughout the network. In summary, AF-NET combines advanced encoding techniques with attentive feature fusion to optimize feature extraction and improve the accuracy of

pixel-level predictions in image processing tasks, making it particularly suited for applications such as segmentation in remote sensing and other complex visual datasets.

6. Impact of the attention mechanism on the skip connection

In this section, we considered the impact of cSAM, cCAM, sfSCAM, and pfSCAM on the skip connection of UNet. At the end of this section, we considered the applications, which are solved by various attention mechanisms using their impact on the encoder of UNet (Table 2). We also discussed the image modality and loss functions used in applications.

6.1. Impact of convolution-based spatial attention module on the skip connection of UNet

Oktay et al. [23] introduced a novel attention gate (AG) inspired by previous work [84], incorporating a grid-based gating mechanism to

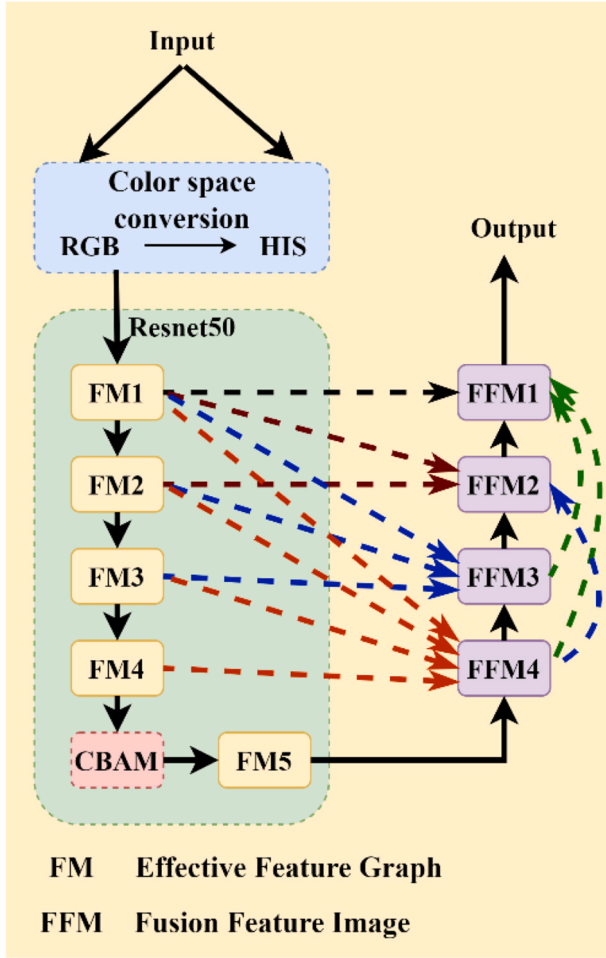


Fig. 11. Impact of sfSCAM on eUNet [51].

enhance the specificity of attention coefficients for local regions. These attention gates were integrated into the basic UNet architecture to emphasize salient features propagated through skip connection. Specifically, in their approach, FMs were extracted at a coarse scale and were utilized within the gating mechanism. This application aimed to filter out irrelevant and noisy responses within the skip connections. Importantly, this gating operation occurred immediately before concatenation operations, ensuring that only pertinent activations were merged.

By leveraging these attention gates, Oktay et al. [23] enhanced the UNet architecture's ability to focus on and preserve critical features during upsampling, thereby improving the accuracy and efficiency of tasks such as image segmentation and reconstruction in medical imaging and other visual recognition applications. Layer by layer, Convolution layers use local information to gradually extract higher-dimensional feature representations (x^l). At the output layer l , the feature map x^l was produced by progressively applying a non-linear activation function and a linear transformation. The ReLU: $\delta_1(x_{i,c}^l) = \max(0, x_{i,c}^l)$ was frequently selected, with I and c for representing the spatial and channel dimensions, respectively. $x_c^l = \delta_1(\sum_{c' \in F_1} x_{i,c'}^{l-1} * K_{c',c})$ is the formula for feature activation, where $*$ stands for the convolution operation.

By identifying prominent regions and pruning feature responses, the attention coefficient $\beta_l \in [0, 1]$ maintained only those activations that were pertinent to a given task. The input feature map with attention coefficient was multiplied element-wise by the attention gate AGs, yielding the following output: $\hat{x}_{i,c}^l = x_{i,c}^l \cdot \beta_l$. Under the default configuration, each vector $x_i^l \in S_{G_l}$, where G_l is the number of feature-maps at

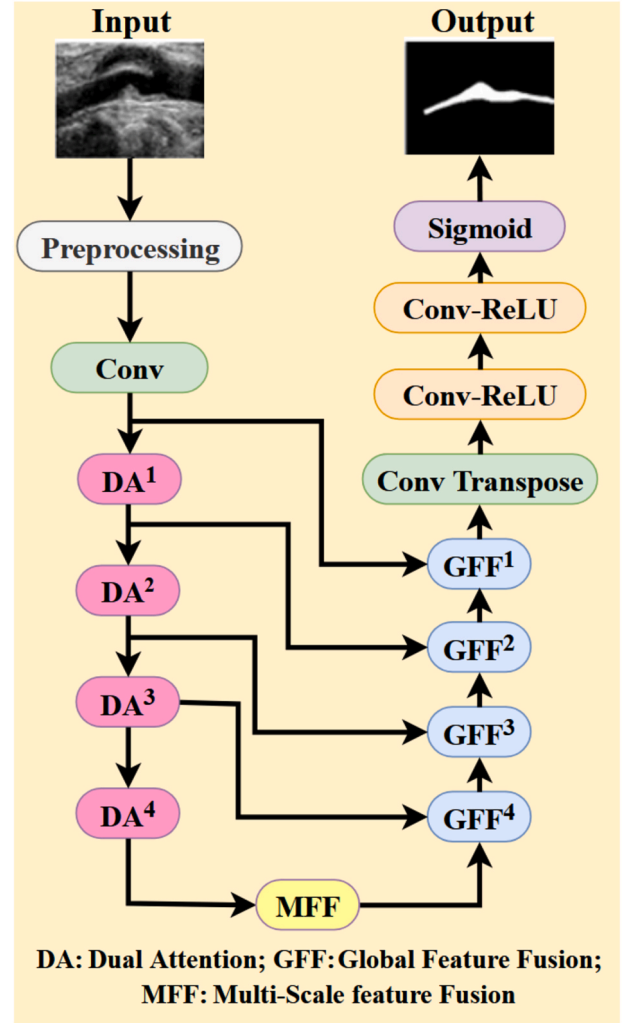


Fig. 12. (a) Impact of pfSCAM on eUNet (b) DA-based encoder (c) MFF and GFF-based decoder block [26].

layer l , had a single attention value computed for it. Inspired by Ref. [85], the author suggested learning multi-dimensional attention coefficients in cases with several semantic classes. To ascertain the focus regions for each pixel, I , a vector $h_i \in S^G$ was employed. As mentioned in Ref. [86], the gating vector uses contextual information to traverse the replies based on lower-level features. The authors used additive attention [87] to get the attention coefficient. Additive attention is more computationally costly than multiplicative attention [88], but it achieves higher accuracy. Formulating additive attention as:

$$r_{att}^l = \Theta^T(\delta_1(Y_x^T x_i^l + Y_h^T h_i + c_h)) + c_\emptyset \quad (11)$$

$$\beta_i^l = \Theta_2(r_{att}^l(x_i^l, h_i; \theta_{att})) \quad (12)$$

The sigmoid activation function is represented by $\Theta_2 = \frac{1}{1 + \exp(-x_{i,c})}$. A set of parameters called θ_{att} , which include linear transformations, define AG. $c_\emptyset \in S$, $c_h \in S^{G_{int}}$, and bias terms containing linear transformations: $Y_x \in S^{G_1 \times G_{int}}$, $Y_h \in S^{G_h \times G_{int}}$, $\emptyset \in S^{G_{int} \times 1}$ and $c_\emptyset \in S$, $c_h \in S^{G_{int}}$. For input tensors, channel-wise $1 \times 1 \times 1$ convolutions are used to compute the linear transformations. Fig. 15 illustrates the impact of cSAM on sUNet, highlighting how the integration of SCAM refines feature representations by focusing on spatial and channel-wise information. This enhancement further improves the specificity and quality of the features propagated through the network, ultimately leading to better performance in image segmentation task.

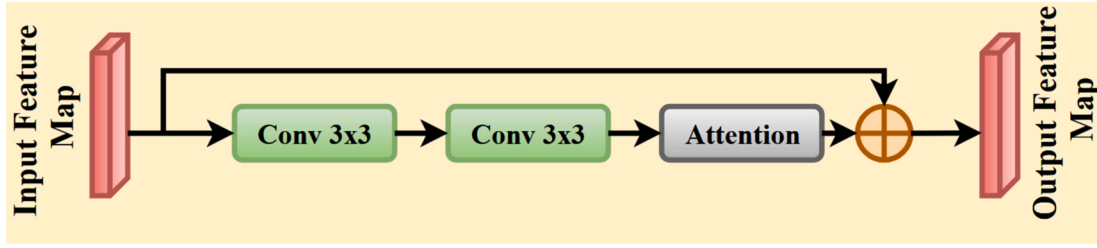


Fig. 13. Architecture of DA block [26].

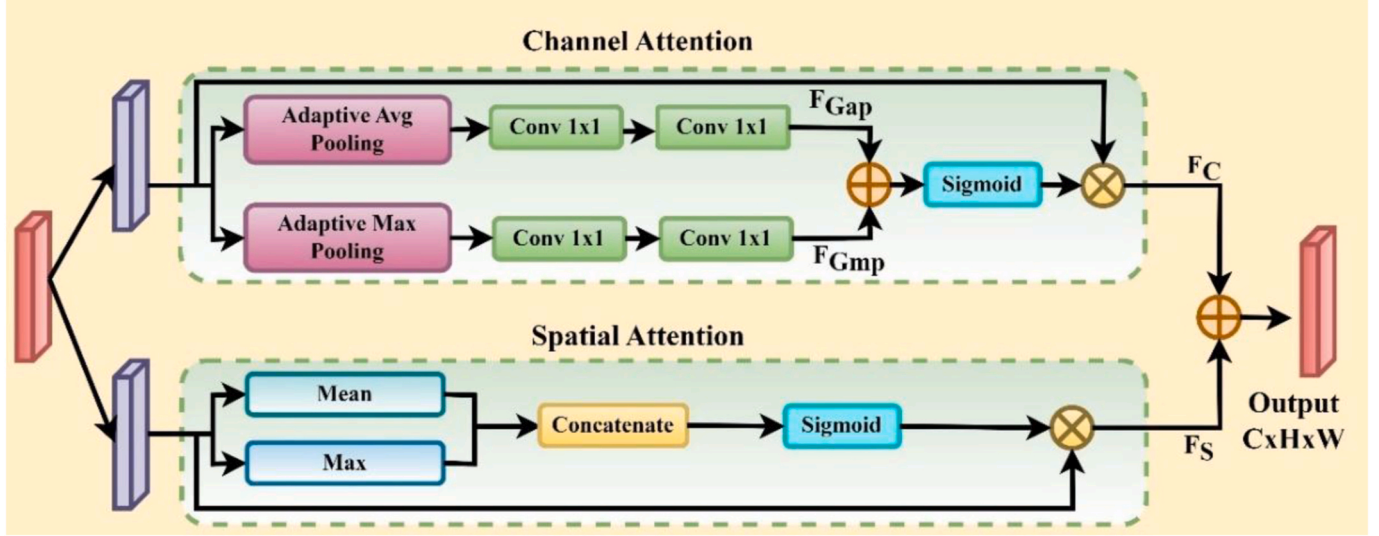


Fig. 14. Architecture of pfSCAM [26].

6.2. Impact of convolution-based channel attention module on the skip connection of UNet

Wang et al. [89] introduced HAD-ResUNet for medical image segmentation, incorporating long-short skip connections to better combine low- and high-level features. The model replaces the UNet bottleneck with hybrid dilated convolutions to capture multi-scale structures and contextual information more effectively. A key component of the architecture is the convolution-based channel attention module (cCAM), which enhances channels carrying critical semantic cues while suppressing those contributing little to the prediction. This combination of hybrid dilation and channel attention improves both local detail preservation and global context understanding, resulting in higher segmentation robustness in complex medical settings.

Since each channel encodes a distinct semantic representation, the authors added a channel attention (CA) block to every skip connection to remove redundant feature maps and retain only task-relevant information. For an input feature map $I \in \mathbb{R}^{C \times H \times W}$, the CA block first reshapes I into matrices K and $Q \in \mathbb{R}^{(H \times W) \times C}$. Next, the product QK is scaled by $\sqrt{C}$, followed by a SoftMax operation to generate the cCAM matrix $A \in \mathbb{R}^{C \times C}$, as shown in Fig. 16. This attention matrix captures channel-to-channel dependencies and allows the network to emphasize semantically important channels during feature fusion.

$$a_{ji} = \Theta \left(f(I_i, I_j) \sqrt{C} \right) \quad (13)$$

Where Θ denotes the sigmoid function, a_{ji} reflects the influence of the i -th channel on the j -th channel, and the function f models their pairwise relationship. Global Average Pooling (GAP) computes channel-wise statistics from the feature map $I \in \mathbb{R}^{C \times H \times W}$, producing an attention

vector as in Equation (14). This vector captures global semantic information and performs dimensional compression while preserving salient content. After obtaining the attention vector A , the authors multiplied it with V and reshaped the result to $\mathbb{R}^{C \times 1 \times 1}$. This output was then scaled by a learnable parameter γ and added to I following a 1×1 convolution, yielding the final output $O \in \mathbb{R}^{C \times H \times W}$. This process incorporates global context into the feature map, strengthens important channel responses, and improves overall feature representation quality.

$$g(I_k) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W I_k(i, j) \quad (14)$$

where g stands for GAP and $k = 1, 2, \dots, C$, $I = (I_1, I_2, \dots, I_C)$.

$$O_j = \gamma \sum_{i=1}^C (a_{ji} \cdot g(I_k) + W_0 I_j + b_0) \quad (15)$$

is the corresponding bias. The final output O is obtained by combining the self-attention output with the convolutional features, weighted by γ . This formulation models channel correlations using spatial information from all feature locations, while global pooling enables channel-wise context aggregation across the entire feature map. Together, these operations strengthen high-level semantic relationships and improve feature selection. Fig. 17 shows how cCAM enhances sUNet by emphasizing informative channels and suppressing irrelevant ones, ultimately improving segmentation performance

6.3. Impact of serial fusion of spatial and channel attention on the skip connection of UNet

Khanh et al. [72] improved the classical UNet by addressing the

Table 2

Segmentation of medical images with the help of an attention mechanism on sUNet. FOV: Field of View, LF: Loss Function, Attn.: Attention. CT: Computed Tomography, MRI: Magnetic Resonance Imaging, US: Ultrasound, pfSCAM: parallel fusion of Spatial and Channel Attention Module, cSAM: convolution-based Spatial Attention Module, cCAM: Convolution Based Channel Attention Module, sfSCAM: serial fusion of Spatial and Channel Attention Module; BCE: Binary Cross Entropy; IoU: Intersection over Union, MSE: Mean Squared Error.

SN	Author	Image Modality	FOV (Application)	Attn. Type	LF Type
1	Oktay et al. [23]	CT	Gastric Cancer/Pancreas	Csam	Dice loss
2	Schlemper et al. [61]	CT/US	Gastric Cancer/Pancreas	cSAM	Dice loss
3	Machireddi et al. [62]	MRI	Malignancy Detection	cSAM	Dice loss
4	Li et al. [63]	CT	Liver Tumor	cSAM	Dice loss/CE
5	Li et al. [64]	X-ray/Satellite/Camera	Skin cancer/Cell nuclei/Coronary artery	cSAM	Dice loss/CE
6	Li et al. [65]	MRI/CT	Liver Tumor/Spleen/Kidney	cSAM	Dice/CE/Focal BCE
7	Wu et al. [66]	Satellite/Camera	Retinal vessel	cSAM	BCE
8	Tang et al. [67]	US	Breast Cancer/Thyroid	cSAM	BCE/Dice
9	Abraham et al. [68]	US	Breast Cancer/Skin Lesion	cSAM	FTL
10	Liu et al. [69]	CT	Liver Tumor/Pancreas/Kidney tumor	pfSCAM	Dice Loss
11	Huang et al. [70]	US	Median Nerve	cSAM	CE
12	Su et al. [71]	CT	Aneurysm segmentation	cSAM	BCE/Dice
13	Khanh et al. [72]	MRI/CT	PolyP/Lung Tumor/Brain Tumor	sfSCAM	Dice loss
14	Cheng et al. [73]	Satellite/Camera	Lung OAR	cSAM	CE
15	Zhang et al. [74]		Nuclei and cytoplasm	cSAM	Dice loss
16	Lin et al. [75]	Dermoscopy	Polyp/skin lesion	cSAM	BCE/IoU
17	Wang et al. [76]	MRI	Brain Tumor	sfSCAM	Per pixel CE
18	Lo et al. [77]	MRI	Whole body fetal MRI	cSAM, cCAM	BCE
19	Chenerlogh et al. [78]	MRI/US	PROSTATE/POLYP/FOOT	cSAM	Dice loss
20	Wang et al. [79]	CT	Liver/Tumor/Neurons	cCAM	BCE
21	Aboussaleh et al. [80]	MRI	Brain Tumor	cSAM	Dice
22	Agarwala et al. [81]	MRI	Brain Tumor	pfSCAM	Dice
23	Liu et al. [82]	Satellite/Camera	Building segmentation	sfSCAM	BCE
24	Cao et al. [83]	Satellite/Camera	Thermal stress	cSAM	MSE

limitations of traditional skip connections, where direct concatenation can pass redundant low-level features and create semantic mismatches. To resolve this, they introduced the serial spatial-channel attention gate (scAG), forming the scAG-UNet architecture (Fig. 18), which includes an encoder, decoder, fusion module, and prediction head. While the encoder extracts multi-scale representations and the decoder recovers spatial structure, scAG refines the information exchanged through each skip pathway.

The scAG integrates a Spatial Attention Gate (sAG) and a Channel Attention Gate (cAG), allowing the network to highlight spatially

discriminative regions and emphasize semantically important channels. The sAG improves localization of fine structures, and the cAG aligns encoder features with decoder semantics, reducing feature inconsistency. This dual-attention design enhances anatomical boundary preservation and supports higher segmentation accuracy in medical imaging. Given encoder features $G_e \in \mathbb{R}^{C_1 \times H \times W}$ and decoder features G_d , scAG performs serial fusion of the spatial attention map $M^S \in \mathbb{R}^{1 \times H \times W}$ and the channel attention map $M^C \in \mathbb{R}^{C_1 \times 1 \times 1}$. The fusion is defined as:

$$G'_e = G_e \otimes M^S(G_e, G_d) \otimes M^C(G_e, G_d) \quad (16)$$

where $\otimes$ denotes element-wise multiplication. Here, M^S is broadcast along the channel dimension, while M^C is extended spatially to match the dimensions of G_e . Fig. 19 depicts the scAG in sUNet, demonstrating improved feature consistency and integration across encoding and decoding stages. Unlike simple skip connections [23], the proposed sAG highlights *critical regions*, directing the network's attention to salient areas (Fig. 20).

To derive spatial attention, the encoder and decoder features are processed with average pooling (AP), max pooling (MP), and 1×1 convolution, each capturing distinct spatial cues. These outputs are concatenated and processed by a 7×7 convolution to build spatial feature maps:

$$M_e^S(G_e) = G_1^{7 \times 7}([\text{AP}(G_e), \text{MP}(G_e), G_1^{1 \times 1}(G_e)]) \quad (17)$$

$$= G_1^{7 \times 7}([(G_e)_{\text{AP}}^s, (G_e)_{\text{MP}}^s, (G_e)_{1 \times 1}^s])$$

$$M_d^S(G_d) = G_1^{7 \times 7}([\text{AP}(G_d), \text{MP}(G_d), G_1^{1 \times 1}(G_d)]) \quad (18)$$

The final spatial attention map is computed by summing encoder- and decoder-derived maps and applying a sigmoid activation:

$$M^S(G_e, G_d) = \Theta(M_e^S(G_e) + M_d^S(G_d)) \quad (19)$$

This module emphasizes spatially important pixels by combining complementary spatial context from the encoder-decoder. Encoder features often have strong spatial resolution but limited semantics, whereas decoder features are semantically rich but spatially coarse. To reconcile this imbalance, cAG strengthens contextual semantics in encoder features. Following the approach in Ref. [26], spatial dimensions are squeezed via AP and MP, resulting in:

$$G_{\text{AP}}^C, G_{\text{MP}}^C \in \mathbb{R}^{C_1 \times 1 \times 1}$$

These descriptors are passed through shared $N^{1 \times 1}$ convolutions, where encoder channels are first reduced to $\frac{C_1}{16}$ to reduce parameters, generating compressed channel dependencies. Summation and sigmoid activation yield the final channel attention:

$$M_e^C(G_e) = G_N^{1 \times 1}(\text{AP}(G_e)) + G_N^{1 \times 1}(\text{MP}(G_e)) \quad (20)$$

$$= G_N^{1 \times 1}((G_e)_{\text{AP}}^c) + G_N^{1 \times 1}((G_e)_{\text{MP}}^c)$$

$$M_d^C(G_d) = G_N^{1 \times 1}(\text{AP}(G_d)) + G_N^{1 \times 1}(\text{MP}(G_d)) \quad (21)$$

$$= G_N^{1 \times 1}((G_d)_{\text{AP}}^c) + G_N^{1 \times 1}((G_d)_{\text{MP}}^c)$$

$$M^C(G_e, G_d) = \Theta(G_{C_1}^{1 \times 1}(M_e^C(G_e) + M_d^C(G_d))) \quad (22)$$

As illustrated in Fig. 21, this mechanism identifies the most informative channels by exploiting inter-channel correlations, thereby enhancing semantic feature transfer and improving decoder guidance. Attention mechanisms within skip connections are particularly effective for small-target segmentation, where precise boundary delineation is critical. By filtering encoder features before fusion, spatial-channel

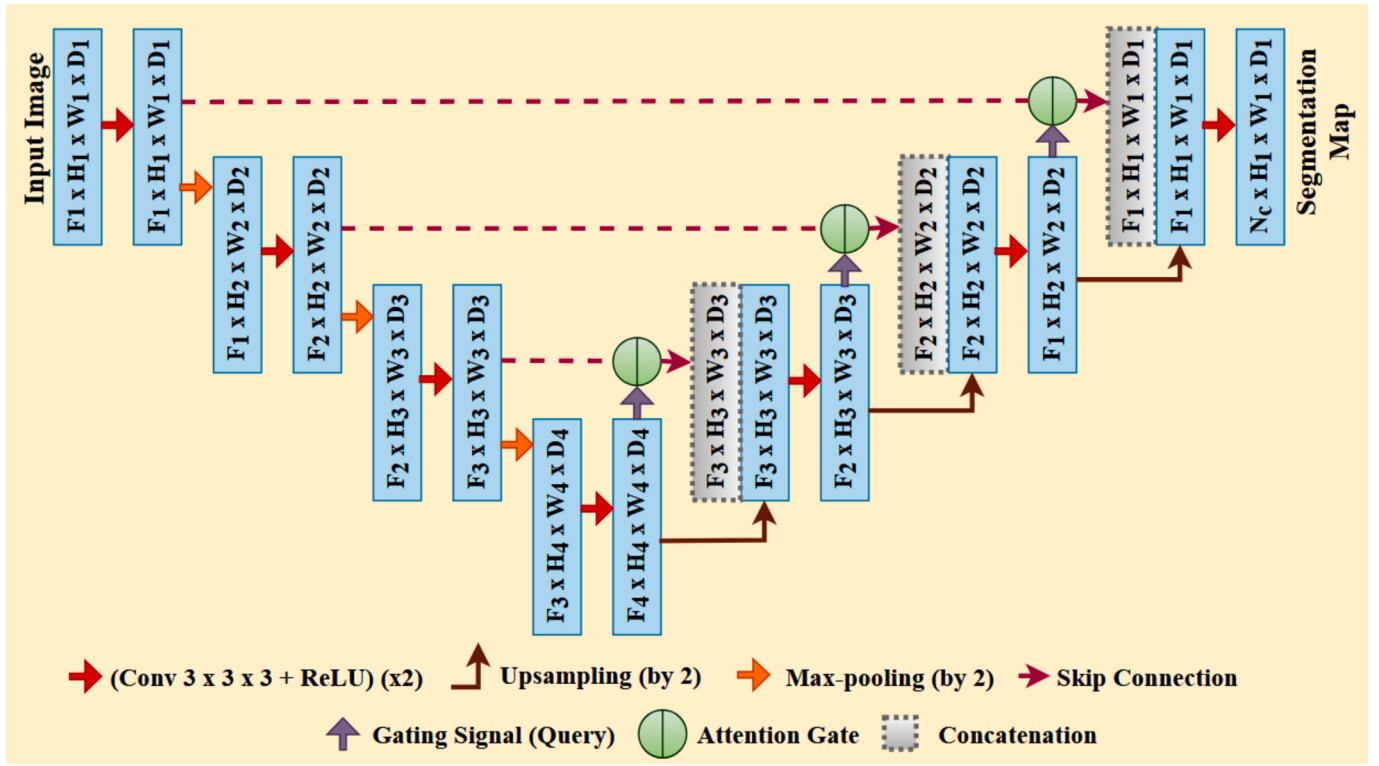


Fig. 15. Impact of cSAM on sUNet [23].

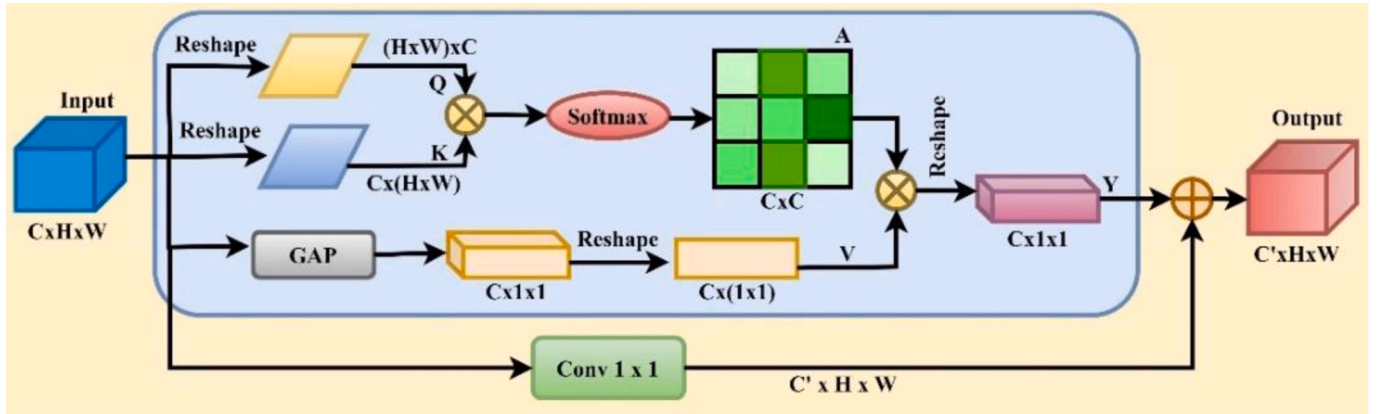


Fig. 16. Architecture of cCAM [89].

attention (e.g., serial fusion) suppresses background noise while preserving weak, discriminative signals. This facilitates improved localization and boundary refinement for small anatomical structures, although most studies report these gains through global metrics rather than boundary-specific evaluations.

6.4. Impact of parallel fusion of spatial and channel attention on the skip connection of UNet

In Fig. 22, the network architecture consists of encoder and decoder modules shown in blue and pink, with dotted lines indicating skip connections. Early layers contain low-level spatial features, so attention gates are introduced to effectively merge information from both the upsampling and downsampling paths. Each attention block appears as a green circle with a plus sign and receives two inputs: one from lower layers carrying higher-level semantic features and one from upper layers providing strong spatial detail, as illustrated in Fig. 23. A $1 \times 1 \times C$

convolution compresses spatial information to generate an $H \times W \times 1$ 3D attention map, while average pooling is performed in parallel to produce a $1 \times 1 \times C$ channel descriptor that is processed by a neural network to compute channel correlations.

7. Impact of the attention mechanism in the decoder of UNet

In this section, we discuss the impact of cSAM, cCAM, sfSCAM, pfSCAM on the decoder of UNet. At the end of the section, the applications that are solved by various attention mechanisms are discussed, using their impact on the encoder of UNet (Table 3). We also discussed the image modality and loss functions used in applications.

7.1. Impact of convolution-based spatial attention module on the decoder of UNet

Fang et al. [92] introduced SCAM-Net (Fig. 27), which comprises

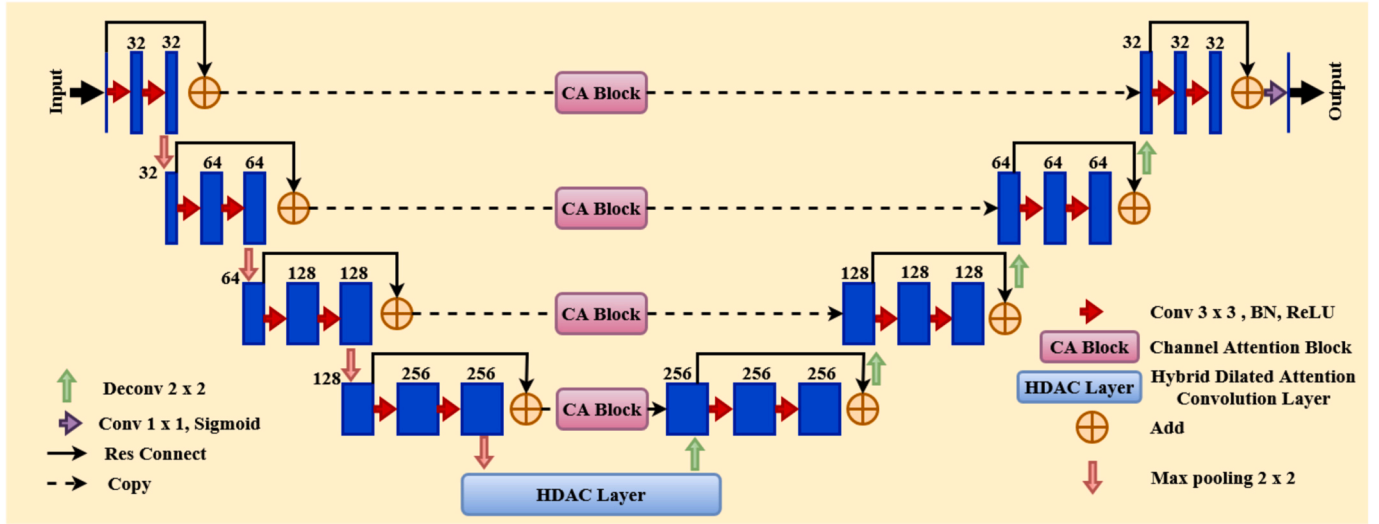


Fig. 17. HAD-ResUNet architecture with cCAM-based channel attention blocks integrated into skip connections to enhance feature fusion and segmentation performance.

four encoder and four decoder stages, each block containing three 3×3 convolution layers with ReLU activation. In the encoder, max pooling (MP) with a 2×2 kernel halves the spatial resolution while doubling the number of channels. In the decoder, the inverse occurs: MP expands the feature map size and halves the channel dimensions. The encoder provides multi-scale features with detailed structural information, while the decoder produces semantically enriched representations. Since not all decoder features contribute meaningfully to the final prediction, the authors incorporated an attention mechanism that suppresses irrelevant regions and highlights important spatial areas and channel maps, thereby improving segmentation performance.

$$\bar{X} = X + f_{AM}(X) \quad (23)$$

The function $f_{AM}(\cdot)$ represents the transformation used to generate the attention map. A simple upsampling operation without skip connections can lead to spatial information loss and blurred structural details. UNet resolves this by concatenating encoder feature maps—rich in spatial detail—with decoder feature maps. However, direct concatenation also introduces redundant low-level features. To address this, the authors proposed cSAM, which suppresses low-value activation regions and reduces redundant feature representations. The structure of the proposed cSAM is shown in Fig. 28. Given a decoder feature map $X \in \mathbb{R}^{W \times H \times C}$, SCAM begins by applying a 1×1 convolution to produce a single-channel output, as defined in Equation (24):

$$X_{SAM} = f_{Conv1 \times 1}(X) \quad (24)$$

The attention map X_{SAM} had identical spatial dimensions as X . After a non-linear transformation, an activation function was applied to produce an attention map ranging from $[0,1]$, with coefficients close to 1, highlighting more relevant regions, as represented by the following Equation (25):

$$A_{SAM} = \Theta(X_{SAM}) \quad (25)$$

here, $\Theta(\cdot)$ denotes the sigmoid activation function. The extracted FM was then added to the raw FM X to emphasize more pertinent regions, as shown in Equation (26). In this context, $f_{Ext}^{Ch}(\cdot)$ Indicates the use of spatial attention within the channel dimension, resizing to correspond with X , facilitating its combination with the raw FM before being processed through the main network.

$$\bar{X}_{SAM} = X \odot f_{SAM}(X) = X \odot f_{Ext}^{Ch}(A_{SAM}) \quad (26)$$

7.2. Impact of convolution-based channel attention module on the decoder of UNet

The core idea of cCAM is to automatically determine the indexed weights for each channel map. In this method, maps with higher relevance to the final prediction results are assigned greater importance, whereas those with lower relevance are given less importance. The authors applied channel attention by leveraging relationships among the various channels in the maps shown in Fig. 29. The resulting maps from the decoder block, referred to as X , were combined to form a channel contribution index using GAP, as demonstrated in Equation (27):

$$m_k = \frac{1}{W \times H} \sum_{w=1}^W \sum_{h=1}^H x_k(w, h) \quad (27)$$

where $x_k(w, h)$ denotes the feature value on spatial position w, h , and the channel k of FM in X , and m_k denotes the global information of the k th channel of the FM. In the next step, channel-wise dependencies were investigated by two fully connected (FC) layers. First, FC layers encode the channel's global vector $m = [m_1, m_2, \dots, m_k]^T$ to a vector with reduced dimensions based on a reduction ratio. The subsequent FC layer then restored it to the original channel count k , forming a channel attention vector, as illustrated by the following Equation (28):

$$X_{CAM} = W_2 (W_1 m) \quad (28)$$

where $W_1 \in \mathbb{R}^{K \times K}$ and W_2 signify the parameters involved in both FC layers, and r ($r = 16$ in this paper) represents the ratio of scaling parameters. In the next step sigmoid function was applied to generate an attention map, which was expressed in the given Equation (29):

$$A_{CAM} = \Theta(X_{CAM}) \quad (29)$$

where $\Theta(\cdot)$ Represented as a sigmoid function. Finally, the channel attention modulated FM is calculated as shown in Equation (30):

$$A_{SCAM} = f_{Ext}^{Ch}(A_{Spatial}) + f_{Ext}^{Spa}(A_{Spectral}) \quad (30)$$

where $f_{Ext}^{Spa}(\cdot)$ Illustrated the channel attention map along the spatial dimension, maintaining the dimensions X , and then processed it through the mainstream.

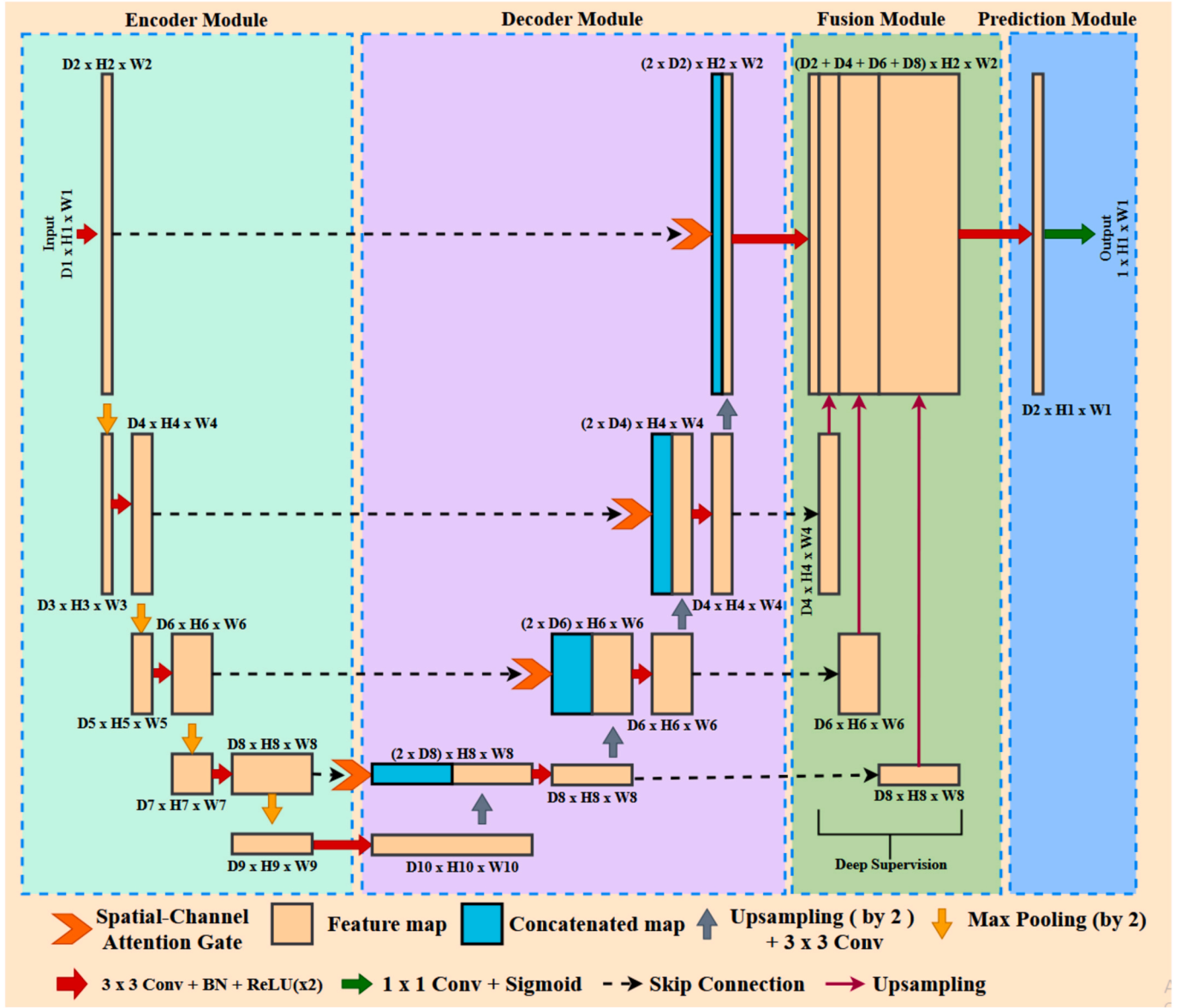


Fig. 18. Impact of sfSCAM on sUNet [72].

7.3. Impact of serial fusion of spatial and channel attention module on the decoder of UNet

Pei et al. [95] proposed a multi-attention-based double UNet for optic cup and disc segmentation, shown in Fig. 24. The cCBAM [25] module was added in the first UNet to make the network focused on spatial and channel features. The T-LSTM module was added to the second UNet. The T-LSTM module contained a double LSTM to learn vertical and horizontal features: first, multi-dimensional input tensors were compressed into two-dimensional features; then, the features were fused with the output of the convolutional layer to obtain globally related features.

The convolutional block attention module (cCBAM) combined cCAM and cSAM. In the first step, attention maps were calculated separately in channel and space. After that, they were multiplied by input features to perform feature refinement, and MP and AP were used to calculate cCAM for each channel of input features. Pooling results were fed into a multi-layer perceptron to calculate cCAM, as shown in Fig. 4. To calculate the cSAM, first the AP and MP operations are applied to the input FM, and then a convolution with a 7×7 kernel is performed.

Finally, these results are passed through the sigmoid function as described in the equation below. (See in Figs. 25–27)

7.4. Impact of pfSCAM on dUNet

As described in the spatial and channel attention module SCAM in sections (4.9 and 4.10) by Fang et al. [92]. The authors combined the SCAM and generated SCAM, which focused more on relevant regions than on individual attention. The authors proposed two aggregation strategies, as depicted in Figs. 28 and 29. The initial strategy, referred to as attention fusion-based SCAM, also known as SCAM-AF, as depicted in Fig. 29, integrated SCAM through element-wise addition, as shown in the following equation:

$$A_{SCAM} = f_{Ext}^{Ch} (A_{Spatial}) + f_{Ext}^{Spa} (A_{Spectral}) \quad (31)$$

Afterward, the attention map was combined with the original FM, resulting in the attention-enhanced feature representation:

$$\bar{X}_{SCAM} = X \odot A_{SCAM} \quad (32)$$

The second technique, called attention-modulated feature fusion

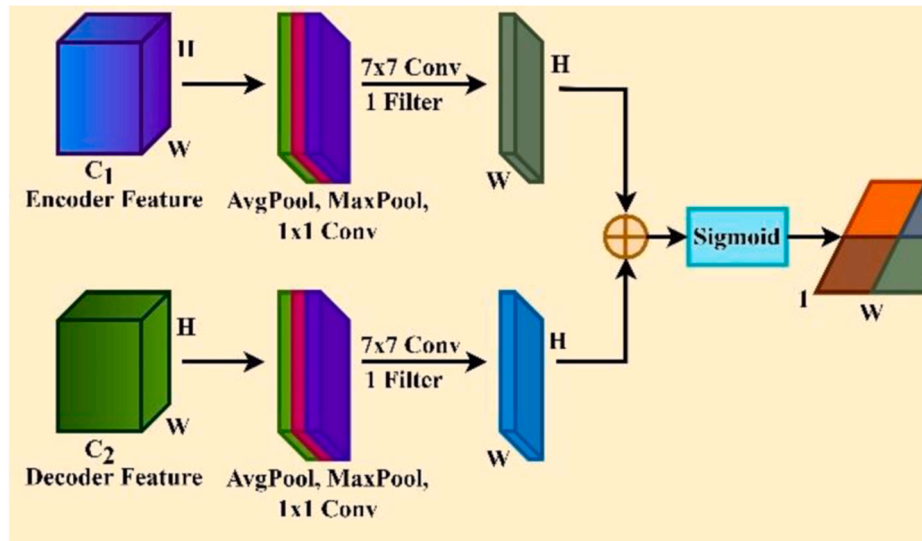


Fig. 19. Architecture of cSAM [72].

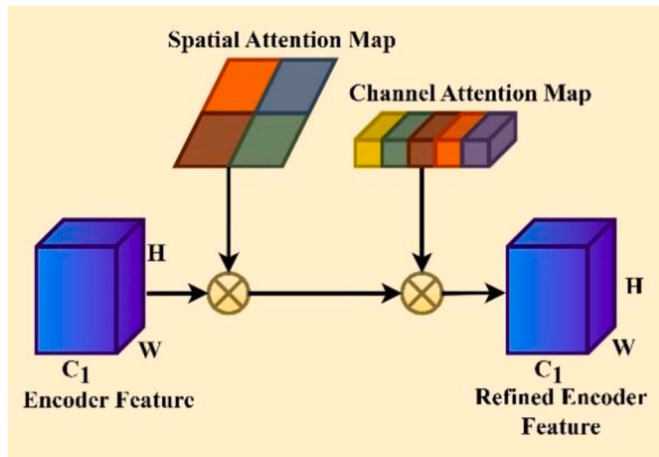


Fig. 20. Serial fusion of cSAM and cCAM [72].

(SCAM-AMFF), combines spatial and channel attention. This method merges independently modulated FMs from SAM and CAM through concatenation, as depicted in Fig. 29, and is described by the following

equation:

$$\bar{X}_{SCAM} = \odot (\bar{X}_{SAM}, \bar{X}_{SCAM}) \quad (33)$$

8. Impact of the attention mechanism in the bridge of UNet

In this section, discuss the impact of cSAM, cCAM, sfSCAM, and pfSCAM on the bridge of UNet. At the end of this section. The applications that are solved by various attention mechanisms using their impact on the encoder of UNet are discussed (Table 4). We also discussed the image modality and loss functions used in applications.

8.1. Impact of convolution-based spatial attention module on the bridge of UNet

Retinal blood vessels are extremely small and fragile, making them difficult to distinguish from the background due to minimal contrast. The difference between vascular and non-vascular features in the retinal fundus image is clear because of the small and marginal vascular areas. To solve this problem, Guo et al. [112] proposed a spatial attention-based SA-UNet for retinal vessel segmentation, as shown in Fig. 30. Spatial attention enhanced important features, such as vascular features, and suppressed irrelevant features, thereby improving

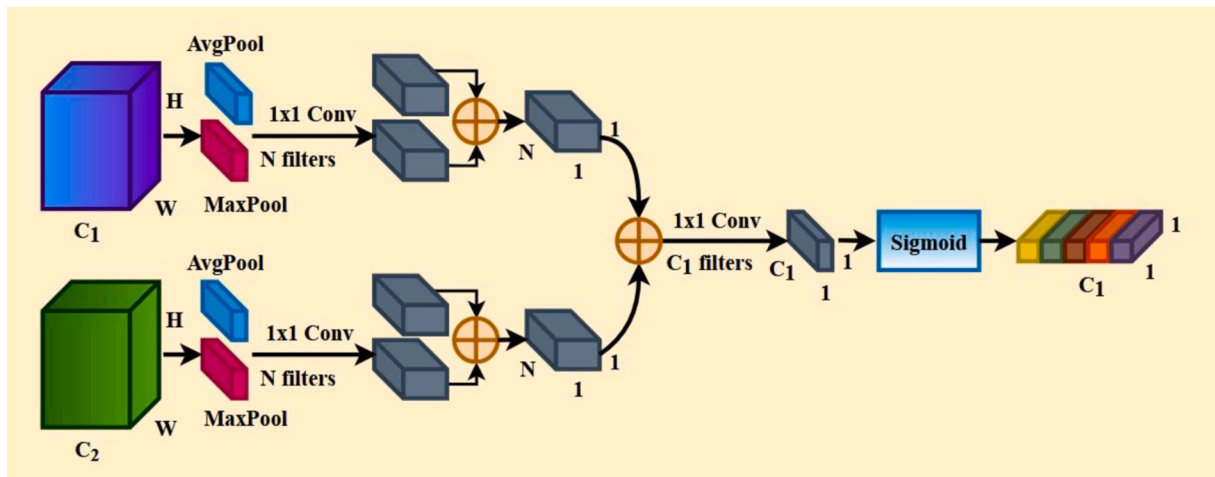


Fig. 21. Architecture of cCAM [72].

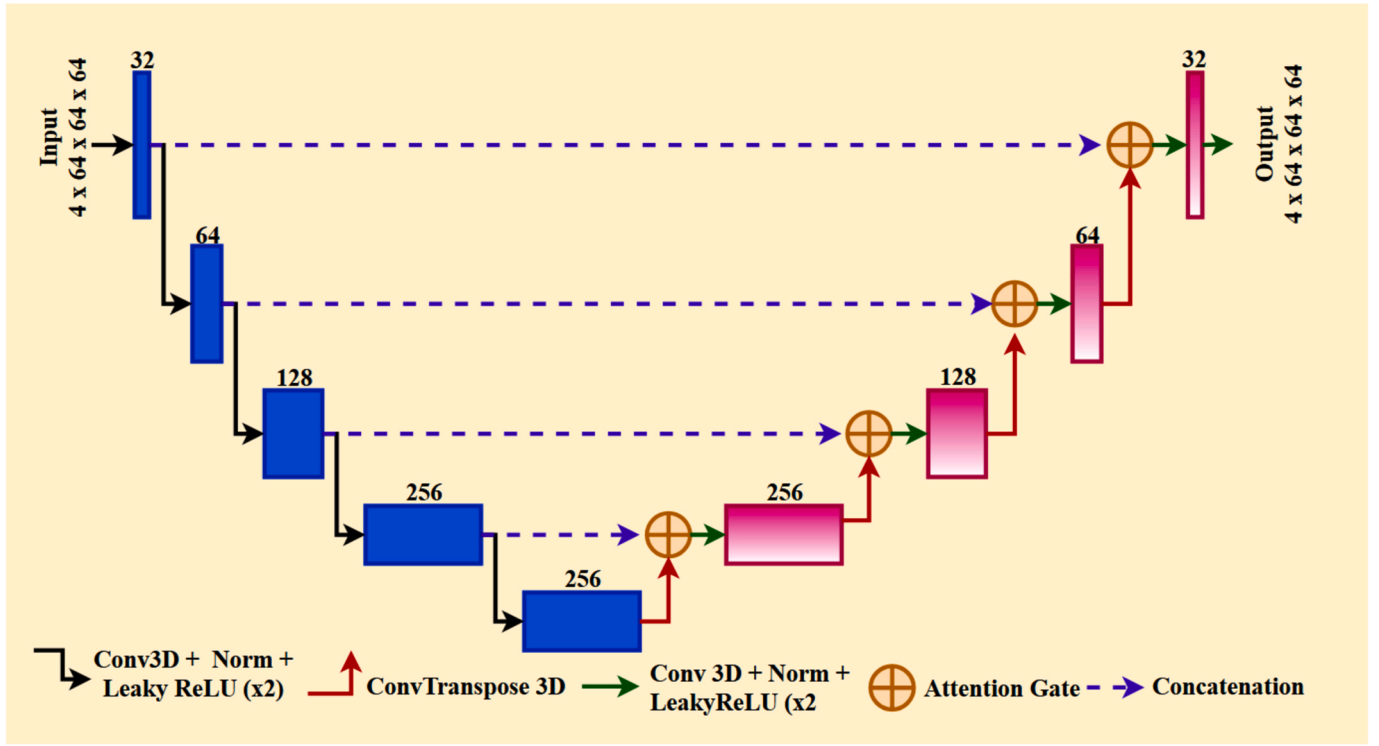


Fig. 22. Impact of pfSCAM on sUNet [81].

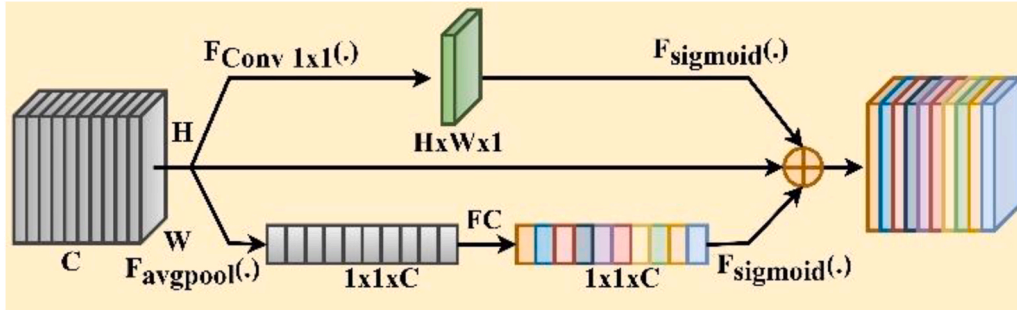


Fig. 23. Parallel fusion of cSAM and cCAM [81].

performance. To calculate spatial attention, first apply the AP and MP operations along the channel axis; afterward, the results are concatenated to produce the feature descriptor shown in Fig. 31. Let us suppose that channel-wise AP and MP operations were applied on input FM $G \in \mathbb{R}^{H \times W \times C}$ and produced output $G_{AP}^S \in \mathbb{R}^{H \times W \times 1}$ and $G_{MP}^S \in \mathbb{R}^{H \times W \times 1}$, respectively. After that, these outputs are concatenated and applied to a 7×7 convolution operation followed by a sigmoid function. The whole process is described in the following equation:

$$\begin{aligned} G^S &= G \cdot M^S(F) \\ G^S &= G \cdot \Theta(g^{7 \times 7}([MP(G); AP(G)])) \\ G^S &= G \cdot \Theta(g^{7 \times 7}([G_{MP}^S; G_{AP}^S])) \end{aligned} \quad (34)$$

8.2. Impact of convolution-based channel attention module on the bridge of UNet

Delibasogulu et al. [112] proposed INCSA-UNET to address the problem of object boundary extraction, as shown in Fig. 32. The architecture contained inception block, drop block, spatial attention, and

channel attention. In the proposed INCSA-UNET, the encoder consisted of five classical sequential layers and four parallel inception layers. If y_i be the input FM of the i^{th} layer, then the output of the classical sequential layer at the encoder side is formulated as:

$$Y_{i+1} = MP(\beta(\beta(y_i g^{3 \times 3}) g^{3 \times 3}) g^{DB}) \quad (35)$$

where $g^{n \times n}$ denotes the filters with $n \times n$ kernel, β denotes ReLU, g^{BN} denotes the BN, and g^{DB} denotes the drop block. The output of the inception layer on the encoder side was obtained by concatenation of the output of different filters used in the proposed inception block (y_a, y_b, y_c, y_d) as shown in Equations (36)–(39):

$$y_a = \beta(\beta(y_i g^{3 \times 3}) g^{DB} g^{BN} g^{3 \times 3}) g^{DB} g^{BN} \quad (36)$$

$$y_b = \beta(\beta(y_i g^{5 \times 5}) g^{DB} g^{BN} g^{5 \times 5}) g^{DB} g^{BN} \quad (37)$$

$$y_c = \beta(\beta(y_i g^{3 \times 3}) g^{DB} g^{BN} g^{1 \times 1}) g^{DB} g^{BN} \quad (38)$$

$$y_d = \beta(y_i g^{1 \times 1}) g^{DB} g^{BN} \quad (39)$$

Table 3

Segmentation of medical images with the help of an attention mechanism on diet. FOV: Field of View, LF: Loss Function, Attn.: Attention. CT: Computed Tomography, MRI: Magnetic Resonance Imaging, US: Ultrasound, pfSCAM: parallel fusion of Spatial and Channel Attention Module, cSAM: convolution-based Spatial Attention Module, cCAM: Convolution Based Channel Attention Module, sfSCAM: serial fusion of Spatial and Channel Attention Module; BCE: Binary Cross Entropy; IoU: Intersection over Union, MSE: Mean Squared Error, NLLLoss: negative loss, GDL: Gradient Difference Loss.

SN	Author	Image Modality	FOV (Application)	Attn. Type	LF Type
1	Ding et al. [90]	MRI/CT	Myocardium/ Great vessels	cCAM	CE/Dice
2	Pennisi et al. [91]	Dermoscopy	Skin Cancer/ Melanoma	cSAM	CE/FTL
3	Fang et al. [92]	CT/ Dermoscopy	Lung/skin	pfSCAM	N
4	Lin et al. [93]	Satellite/ Camera	Industrial defect segmentation	cSAM	BCE
5	Zhao et al. [94]	Satellite/ Camera	Forward-looking Sonar for underwater Glaucoma	cCAM	Dice Loss
6	Chen et al. [95]	CT	Spine	sfSCAM	Dice Loss
7	Wang et al. [96]	CT	Head and neck	sfSCAM	Dice Loss
9	Zhong et al. [97]	MRI/CT	Skin lesion/ Polyp/Melanoma lesion	cSAM	BCE
10	Liu et al. [98]	Dermoscopy	Nuclei Segmentation/ cancer	cCAM	Focal loss
11	Zeng et al. [99]	MRI	Brain tumor	cCAM	Dice/CE
12	Huang et al. [100]	MRI	change detection in satellite	sfSCAM, pfSCAM	Negative log
14	Adil et al. [101]	Satellite/ Camera	Satellite Imagery	pfSCAM	Dice/CE
15	Zhang et al. [102]	Satellite/ Camera	Road scenes	cCAM	NLLLoss
17	Fan et al. [103]	Satellite/ Camera	Road crack	pfSCAM	N
18	Gao et al. [104]	Satellite/ Camera	Hybridoma cell segmentation	sfSCAM	Focal loss
19	Lu et al. [105]	Satellite/ Camera	Brain tumor	cCAM	GDL/CE
20	Noori et al. [106]	MRI	Skin/breast, and lung cancer	sfSCAM	Dice loss
21	Xu et al. [107]	US	Precipitation	cSAM	MSE/ MAE
22	Xu et al. [108]	Satellite/ Camera	Nocasting	cSAM	Dice/CE
23	Guangzhe et al. [109]	Satellite/ Camera	Image segmentation		

$$y_{l+1}^{\text{Inception}} = [y_a, y_b, y_c, y_d] \quad (40)$$

The authors improved the two-level inception architecture by making use of a drop block (overcome the problem of overfitting) and two types of attention block two-level inception architecture by making use of drop block overcome the problem of overfitting) and two types of attention block, namely spatial and channel (to focus more on relevant region). To calculate the cSAM, firstly AP and MP operations were applied to the input feature $F^{(H,W,C)}$ along the channel axis, followed by a 7x7 convolution and sigmoid activation function as described in the equations below:

$$F_{\text{map}} = \theta([\text{MP}(F), \text{AP}(F)] f^{7 \times 7}) \quad (41)$$

In the last step, the final output is calculated by multiplying the spatial attention and input features, as shown in the equation below:

$$F_s = F_1 \otimes F_{\text{map}} \quad (42)$$

Fig. 33 illustrates the architecture of cSAM, demonstrating how

spatial attention is structured within the INCSA-UNET to enhance focus on relevant spatial features. Let $F_1^{H,W,C}$, and $F_{l-1}^{H,W,C}$ denote the input FM of the l th and $l-1$ st layers. In channel attention, the first gating module is applied on the input FM l th layers, which includes $f^{1 \times 1}$ convolution, BN, and ReLU. The output of the gating module (F_g) was up-sampled and concatenated with an FM of the l th layer filtered with $f^{2 \times 2}$ as shown equation below:

$$F_{\text{con}} = [\Omega(F_g), F_1 f^{2 \times 2}] \quad (43)$$

where Ω is represented as an up-sampling operation. In the next step α , $f^{1 \times 1}$, θ , and Ω operation is applied to capture feature weights (w) and finally output FM (F_s) obtained by multiplying the w and input FM (F_1) as shown below equations:

$$w = \Omega(\theta(\alpha(F_{\text{con}}) \otimes f^{1 \times 1})) \quad (44)$$

$$F_s = (F_1 \otimes w) f^{1 \times 1} f^{\text{BN}} \quad (45)$$

Fig. 34 illustrates the architecture of cCAM, detailing how channel attention selectively emphasizes informative features.

8.3. Impact of serial fusion of spatial and channel attention module on the bridge of UNet

Zhao et al. [113] proposed the UNet-cCBAM architecture, in which the cCBAM architecture [25] was fused with UNet [22]. cCBAM was used to identify relevant FMs across channel and spatial regions in the image, with additional attention paid to the boundary regions of the winter-wheat planning area. cCBAM was fused after every convolution in the contracting path, expansion path, and bottleneck, as shown in the diagram below. After narrowing down the spatial dimension of the FM, the cCAM uses GAP and GMP methods to obtain two aspects of global information. It then feeds this into a shared model consisting of a cMLP and a hidden layer to obtain two features. A sigmoid function (Θ) is used to obtain the channel attention graph N^C after these two characteristics were added. Finally, the channel attention graph and the sums of each element are combined to produce the feature vector. The cCAM is expressed as follows:

$$N^C(F) = \Theta(\text{cMLP}(\text{AP}(F)) + \text{cMLP}(\text{MP}(F))) \quad (46)$$

To obtain the cSAM N^S , the cSAM first applies GMP and GAP based on the feature channels. It then combines these to create valid feature descriptors, convolves them with a 7x7 kernel, and applies a sigmoid function to generate spatial attention features. Additionally, Fig. 35 illustrates the impact of sfSCAM on bUNet. The cSAM is expressed as follows:

$$N^S(F) = \Theta(g^{7 \times 7}([\text{AP}(F), \text{MP}(F)])) \quad (47)$$

$$N^{\text{cCBAM}}(F) = N^C(F) + N^S(F) \quad (48)$$

8.4. Impact of parallel fusion of spatial and channel attention module on the bridge of UNet

Firstly, SCAM overlooks cross-dimensional relationships between spatial positions and channels, as it assigns weights based on catching interrelationships between spatial positions or channels. Second, while vanilla skip-connections increase learning efficiency by allowing duplicate moves of unclear information in the FM, they typically produce blurry image features. To increase the accuracy of liver tumor segmentation, Hong et al. [111] suggested a quartet attention UNet (QAUNet), as illustrated in Fig. 36.

The model adopts four parallel branches to realize the quartet attention mechanism. Two of these branches focus on modeling global communication across spatial positions and channel dimensions

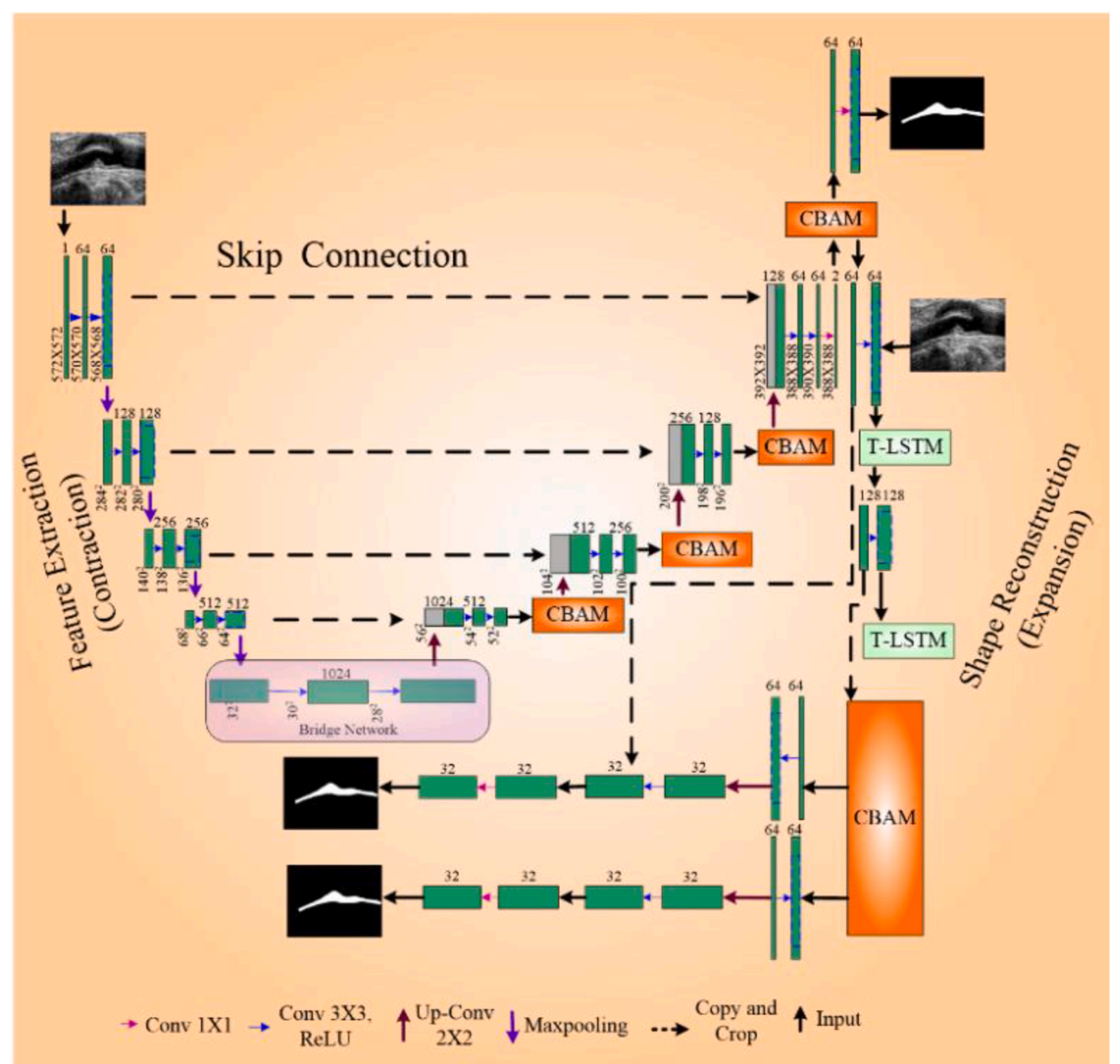


Fig. 24. Impact of sfSCAM on dUNet [95] and convolution block attention module (ccBAM) used as attention block, and LSTM stands for long short-term memory.

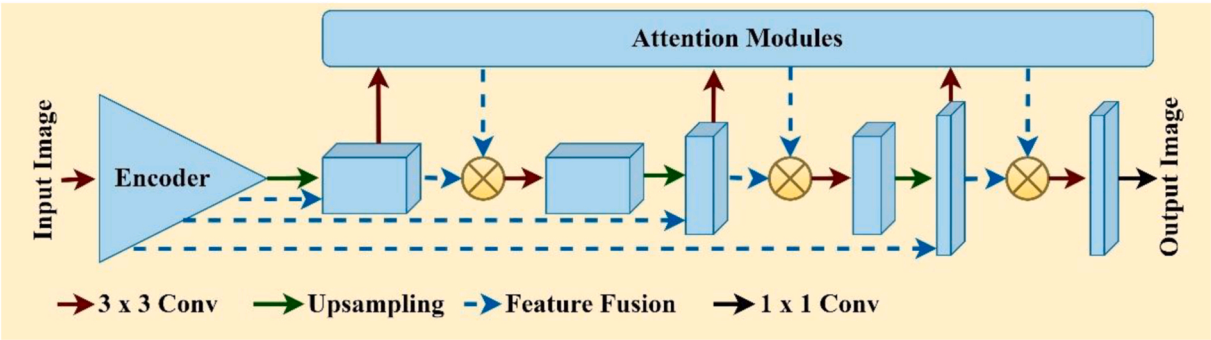


Fig. 25. The architecture of UNet with Attention Module [90].

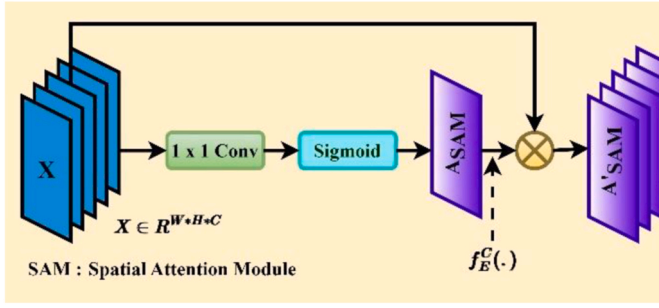


Fig. 26. Architecture of cSAM on dUNet [90].

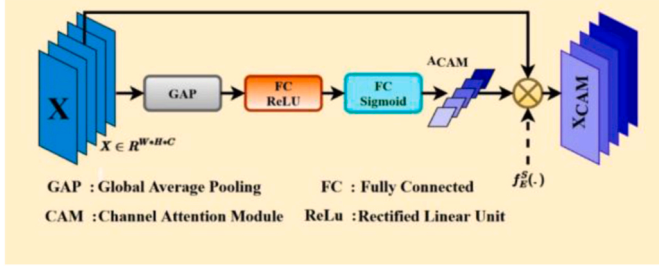


Fig. 27. The architecture of cCAM on diet, global average pooling (GAP), fully connected rectified linear unit (FC ReLU), and fully connected sigmoid (FC sigmoid) used as activation function.

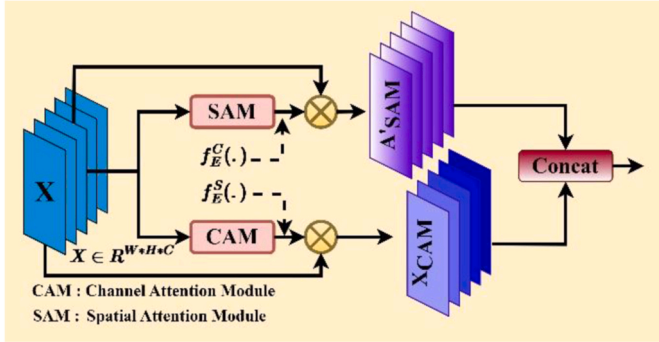


Fig. 28. Architecture of pfSCAM [92] SAM and CAM stand for spatial attention module and channel attention module.

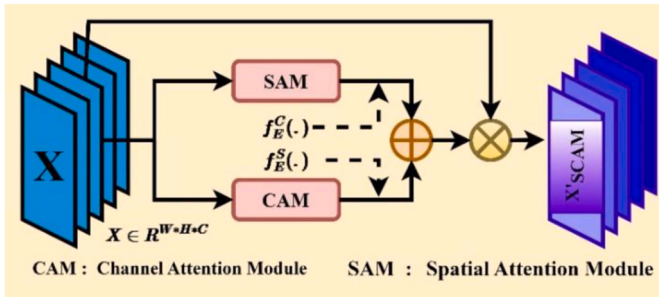


Fig. 29. Architecture of pfSCAM on dUNet [92]. SAM and CAM stand for spatial attention module and channel attention module.

separately. The remaining two branches are responsible for capturing the multidimensional interactions between the channel dimension (C) and the spatial dimensions (H or W). The initial discussion of the QA architecture underscores the significance of cross-dimensional

Table 4

Segmentation of medical images with the help of an attention mechanism on bUNet. FOV: Field of View, LF: Loss Function, Attn.: Attention. CT: Computed Tomography, MRI: Magnetic Resonance Imaging, pfSCAM: parallel fusion of Spatial and Channel Attention Module, cSAM: Convolution-based Spatial Attention Module, cCAM: Convolution-based Channel Attention Module, BCE: Binary Cross Entropy.

SN	Author	Image Modality	FOV (Application)	Attn. Type	LF Type
1	Guo et al. [110]	CT/Satellite/Camera	Retinal Vessel eye	cSAM	BCE
2	Hong et al. [111]	CT	Liver and Liver Tumor	pfSCAM	CE
3	Delibasoglu et al. [112]	Satellite/Camera	Building Segmentation	cSAM, cCAM	Dice loss

interactions. In conventional SCAM, individual attention matrices are constructed for the spatial and channel dimensions and applied to the original feature maps via matrix multiplication to obtain corresponding spatial and channel contextual information.

However, computing spatial and channel attention independently often neglects the intrinsic correlations between these dimensions. To address this limitation, a cross-dimensional interaction mechanism is introduced that explicitly models the relationship between the spatial and channel dimensions within the input tensor. Accordingly, the authors define QA as a four-branch module in which the input and output tensors share the same shape.

Vanilla UNet relies on convolution and pooling operations to progressively reduce spatial resolution and extract deeper semantic representations. Nevertheless, pooling operations frequently obstruct the downward propagation of shallow features, such as edge information. As a result, the network predominantly propagates coarse semantic features, while failing to adequately capture fine edge details or small-scale target features. The following section summarizes the limitations associated with the standard skip-connection design in vanilla UNet:

$$Z^{(l+1)} = Y^{(l-1)} \odot U(Y^l) = Y^{(l-1)} \odot U(C(D(Y^{(l-1)}))) = Y \odot U(Y^l) = (C(Y^{(l-1)} + U(D(Y^{(l-1)})))) \odot U(C(D(Y^{(l-1)}))) \quad (49)$$

To maintain the target's edge features and prevent the down-sampling operation in UNet from obstructing edge information propagation, extra convolutional blocks can be added to filter features that are regularly communicated to the decoder. The following is a concise representation of the long-short skip connection depicted in Fig. 37. Unlike the Vanilla UNet features information on the extra convolution module and the residual path, the module is adaptively filtered by $(C(Y^{(l-1)} + U(D(Y^{(l-1)}))))$.

Fig. 38 displays the QA function, and Fig. 38 shows the QA architecture. Given a $Y \in R^{C \times H \times W}$ as the module's input, here the number of channels is denoted by C, H stands for the height, and W for its width. By defining multiple branches to capture the relationships between the spatial dimensions (H, W) and the channel dimensions (C, W), (H, C), and (C, C) of the input tensor, the authors established cross-dimensional interaction within the Quartet attention mechanism. Adaptive spatial attention is acquired by the first branch. The information in the cSAM is the correlation between pixels, which the authors must gather. The original FM is then multiplied using the cSAM. These two acts capture the long-range contextual information. A simple description of the specific computing procedure is as follows: we first transformed the original input feature, $Y \in R^{C \times H \times W}$, to $Y \in R^{C \times N}$ where $N = H \times W$ is the number of pixels. Next, then compute the cSAM $M \in R^{N \times N}$ by performing a matrix multiplication between the transpose of $Y \in R^{C \times N}$ into $Y_1 \in R^{C \times N}$ and $Y_2 \in R^{C \times N}$. Finally, we apply a softmax layer (see Fig. 39).

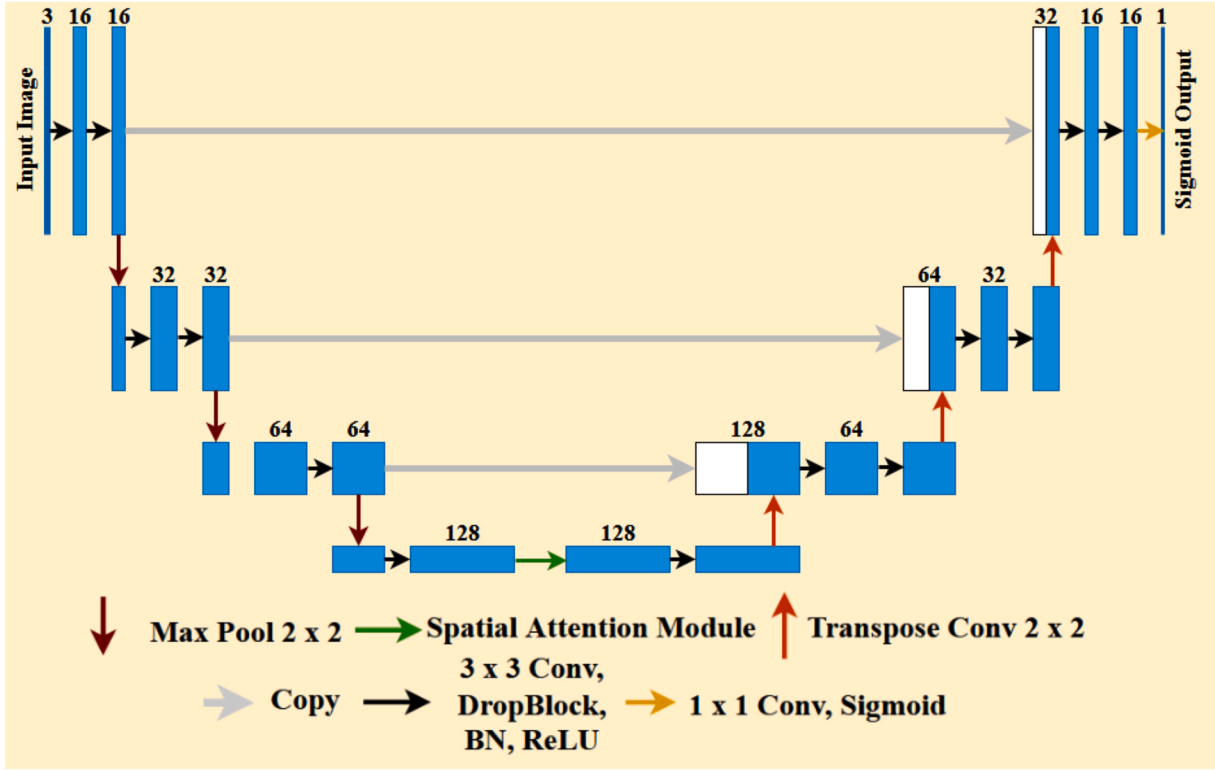


Fig. 30. Impact of cSAM on sUNet [110].

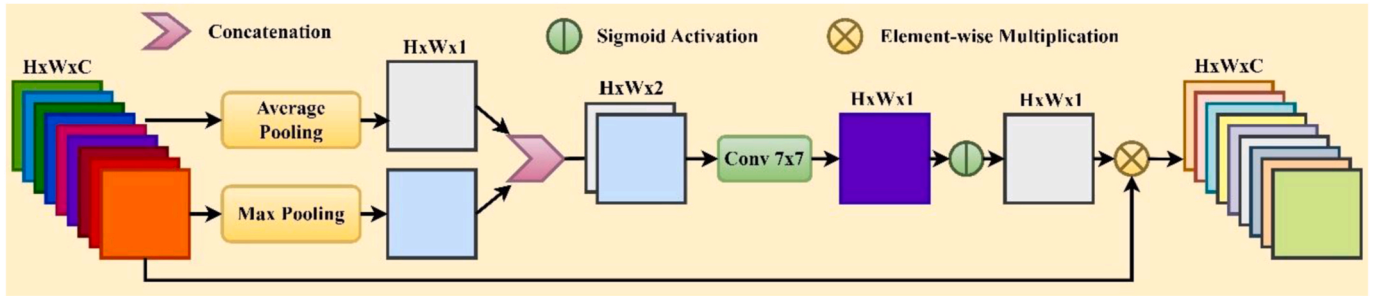


Fig. 31. Architecture of cSAM.

$$m_{ij} = \frac{\exp(Y_{1i} \cdot Y_{2j})}{\sum_{i=1}^N \exp(Y_{1i} \cdot Y_{2j})} \quad (50)$$

where m_{ji} measures how the i -th position affects the j -th position. A greater correlation between the two positions results from their more similar feature representations. The authors simultaneously reshaped $Y \in \mathbb{R}^{C \times H \times W}$ to create a new feature matrix $Q \in \mathbb{R}^{C \times N}$. The result was then reshaped to $\mathbb{R}^{C \times H \times W}$ by the authors using matrix multiplication between Q and M . In order to obtain the final output $out^1 \in \mathbb{R}^{C \times H \times W}$, multiply the result by the scale parameter α , and execute an element-by-element summation using the original feature $X \in \mathbb{R}^{C \times H \times W}$.

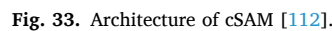
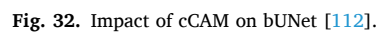
$$Out_j^1 = \alpha \sum_{i=1}^N (m_{ji} Q_i) + Y_j \quad (51)$$

where α starts at 0 and learns to a suitable weight over time. The weighted sum of the features at all points plus the original feature produced the resulting feature out^1 at each position. Like the spatial attention process, the second branch also learned channel attention. The cCAM was computed by the authors by performing a matrix

multiplication between $Y_2 \in \mathbb{R}^{C \times N}$ and $Y_1 \in \mathbb{R}^{C \times N}$ and then applying a softmax layer. $M \in \mathbb{R}^{C \times C}$. We multiply Q by M and transpose $Y \in \mathbb{R}^{C \times H \times W}$ to $Q \in \mathbb{R}^{C \times N}$. We sum over the feature $Y \in \mathbb{R}^{C \times H \times W}$, then multiply the result by the scale parameter α to obtain the final output $out^1 \in \mathbb{R}^{C \times H \times W}$. The long-range semantic relationships between channels are modeled by the produced FM. H and C turned their interactive focus to the third branch. Following the permutation of the input FM, $Y \in \mathbb{R}^{C \times H \times W}$ to $Y_1 \in \mathbb{R}^{C \times H \times W}$, W -dim is used for the GAP and GMP processes, respectively. By taking this step, the computational efficiency is increased, and branch-independent dimension calculation is reduced. After passing through a BN layer and a conventional convolution layer, Y_2 yields an $\mathbb{R}^{1 \times H \times C}$ FM, $Y_2 \in \mathbb{R}^{2 \times H \times C}$. A sigmoid activation layer then produces the resulting attention weights. After applying the attention weights to $Y_1 \in \mathbb{R}^{C \times H \times W}$, the shape is changed to produce output $out^3 \in \mathbb{R}^{C \times H \times W}$. out^3 can be expressed simply as follows:

$$out^3 = \Theta(C(P(Y)))Y_1 \quad (52)$$

where Θ is a convolutional operation, $P(\cdot)$ is a pooling operation, and $\Theta(\cdot)$ is the sigmoid activation. In the same way, C and W 's interactive attention is gained by the fourth branch. The authors permute $Y \in \mathbb{R}^{C \times H \times W}$ to $Y_2 \in \mathbb{R}^{2 \times H \times C}$.



Our systematic analysis reveals that attention mechanisms have emerged as foundational components in the evolution of UNet-based architectures for medical image segmentation. By examining the distribution, taxonomy, and architectural integration of attention modules across more than 800 studies, we show that attention functions not merely as a performance enhancement but as a structural principle shaping modern segmentation models in terms of adaptability,

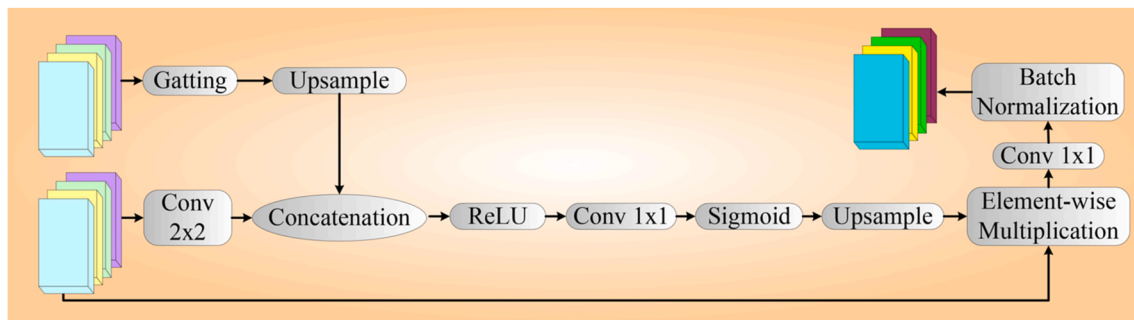


Fig. 34. Architecture of cCAM [112].

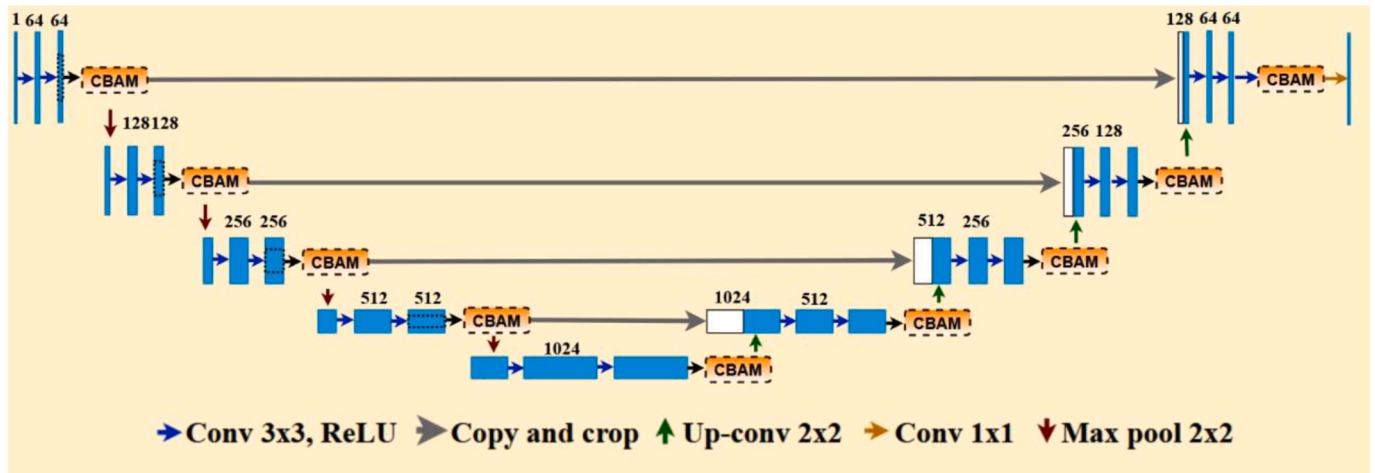


Fig. 35. Impact of sfSCAM on bUNet and cCBAM stands for convolution block attention module.

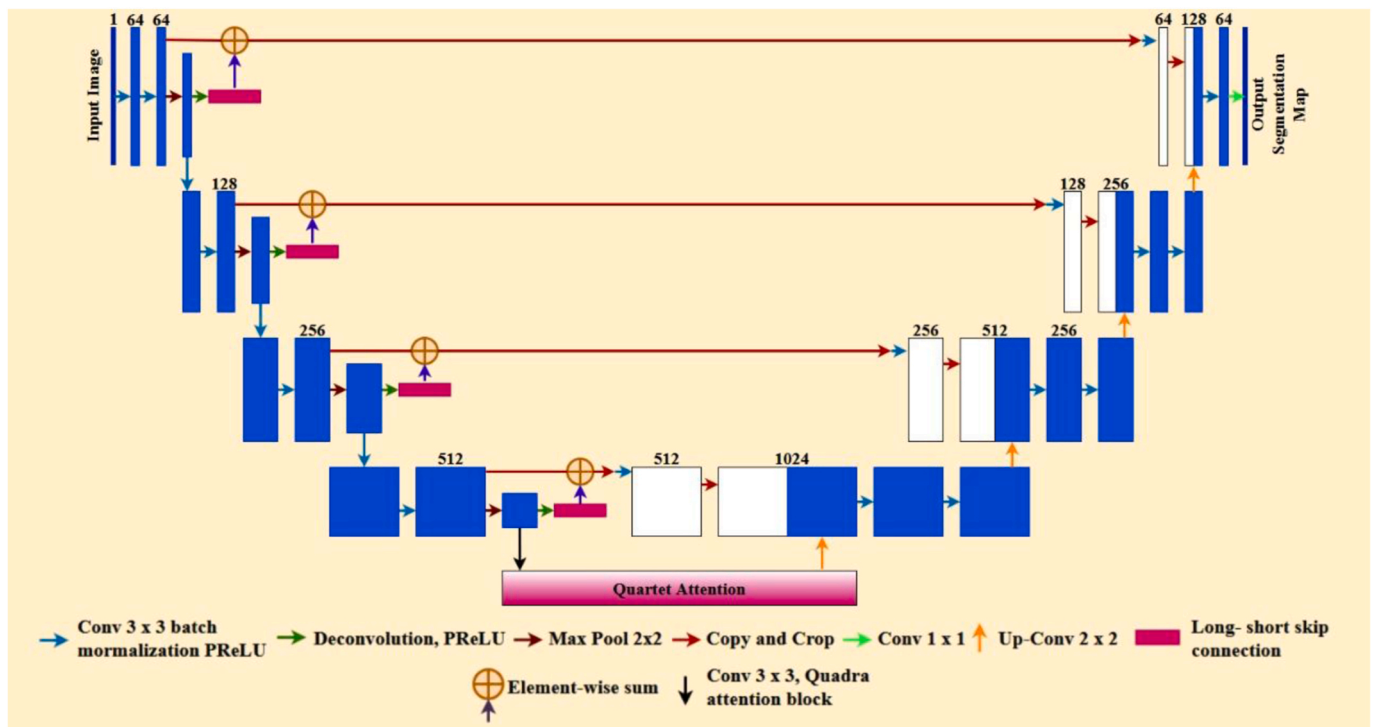


Fig. 36. Quartered attention model.

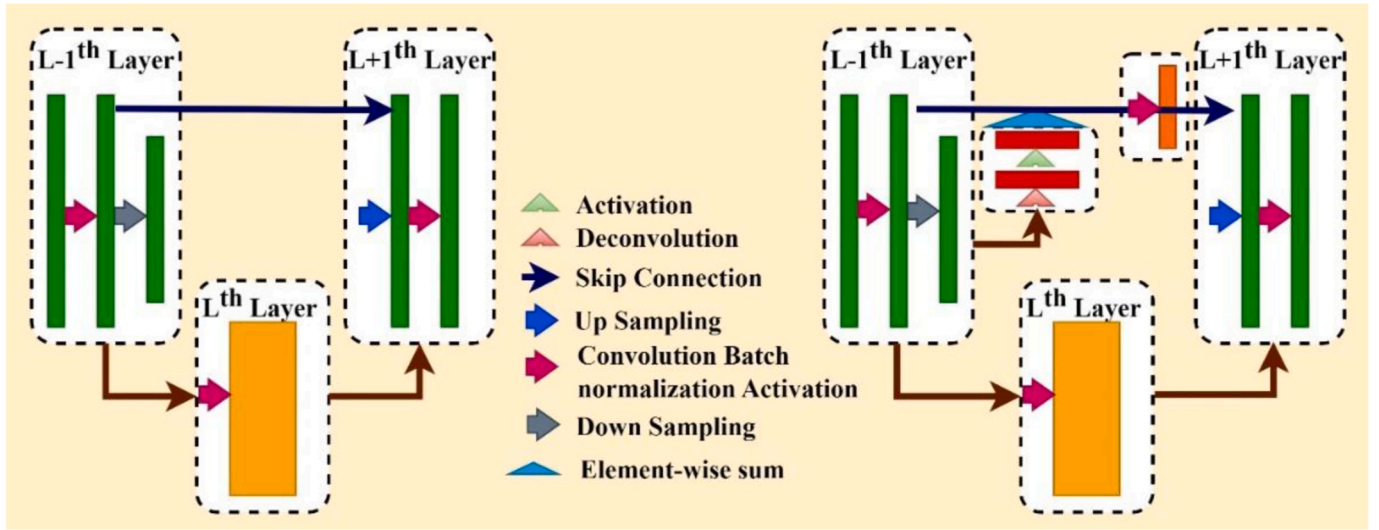


Fig. 37. Architecture of skip-connection and long-short skip-connection.

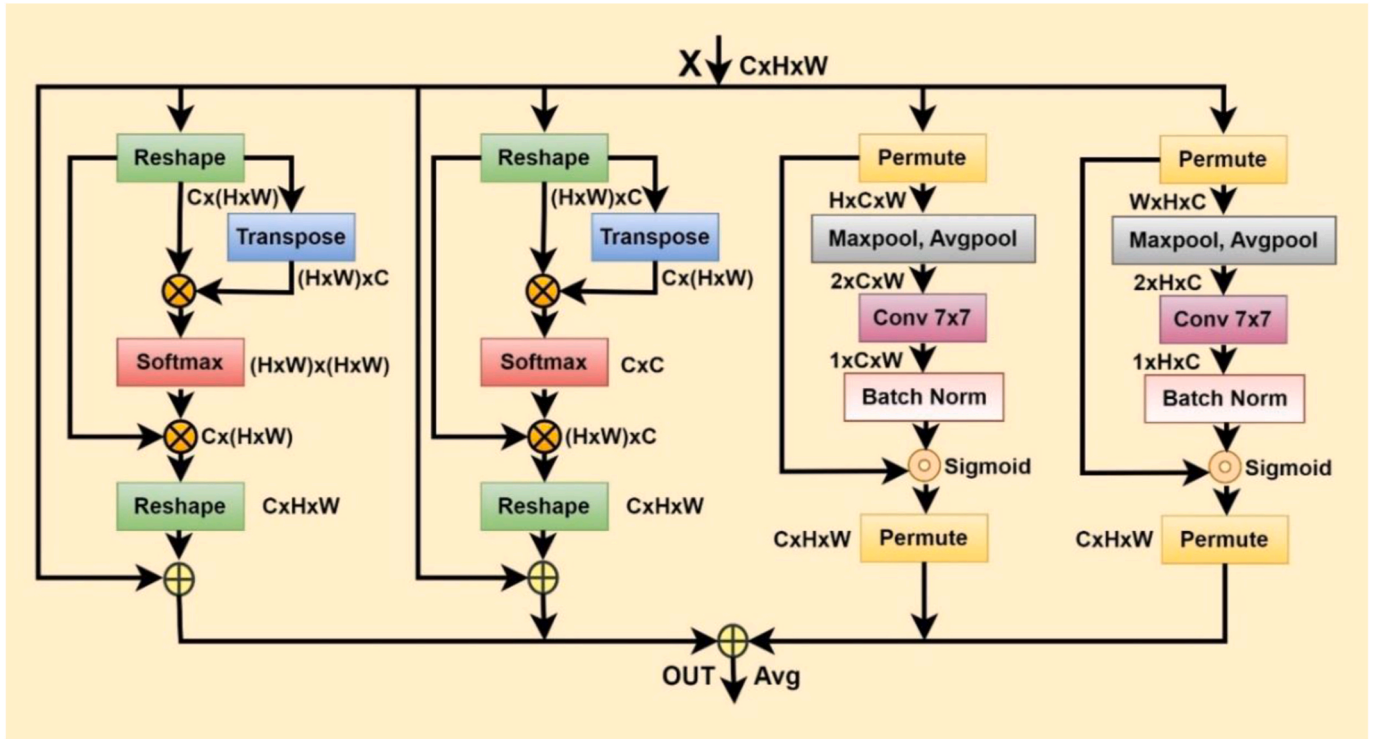


Fig. 38. Computation of the Quartet attention.

interpretability, and robustness. Rather than being confined to a single network stage, attention has become deeply embedded in the design philosophy of contemporary UNet variants.

Attention mechanisms are now systematically embedded across all major components of UNet, including the encoder, decoder, skip connections, bottleneck, and multi-arm configurations (Appendix B)—with each placement contributing distinct and complementary functional improvements (i). Channel attention modules in the encoder selectively amplify semantically salient features, while spatial attention applied within skip connections improves the fidelity of spatial detail transfer between encoding and decoding pathways. Hybrid attention strategies, particularly serial (sfSCAM) and parallel (pfSCAM) fusions of spatial and channel attention, further enhance feature expressivity and robustness

to overfitting by enabling multi-dimensional feature refinement (ii). The sustained dominance of convolution-based spatial (cSAM) and channel (cCAM) attention mechanisms, alongside the rapid rise of transformer-based attention (tAttention) for modeling long-range dependencies and global context, reflects an increasingly deliberate and sophisticated architectural evolution (iii). A PRISMA-based meta-analysis confirms that attention-augmented UNet variants consistently achieve statistically significant performance gains across diverse imaging modalities, including MRI, CT, ultrasound, and fundus photography (iv). These improvements generalize across a wide range of anatomical targets—such as liver and brain tumors, retinal vessels, and skin lesions—demonstrating the broad applicability of attention-driven segmentation models. Attention mechanisms also play a central role in

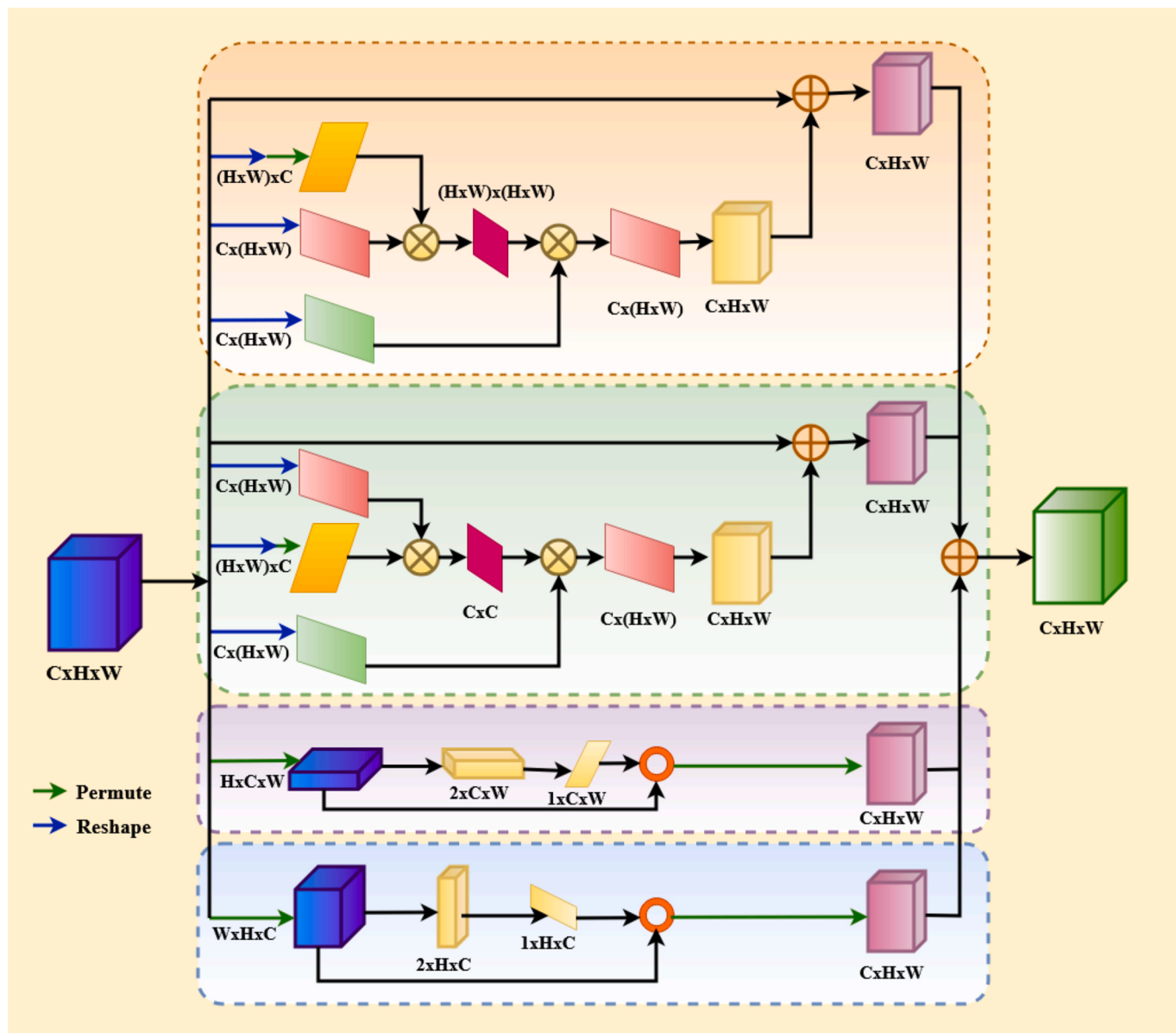


Fig. 39. Model of the Quartet attention.

addressing scale heterogeneity, enabling precise delineation of both fine-grained and large anatomical structures (v). In particular, spatial attention modules such as cSAM, when combined with multi-scale context aggregation techniques like atrous spatial pyramid pooling (ASPP), effectively reconcile multi-resolution feature representations.

An additional trend identified in the literature is the emergence of cascaded and parallel attention pipelines, which reconceptualize attention as a multi-stage refinement process rather than a single operation (vi). Such designs improve feature discrimination while mitigating overfitting and enable substantial architectural flexibility, allowing UNet variants to be deeper, wider, shallower, or pruned for edge deployment without degrading segmentation accuracy. Beyond architectural and performance considerations, attention-based UNet models increasingly support broader AI objectives, including interpretability, robustness under heterogeneous data distributions, and privacy-preserving learning (vii). Their integration into federated learning frameworks and compatibility with explainable AI (XAI) techniques enhance clinical trust by improving feature localization, transparency, and bias mitigation.

Collectively, these findings position attention mechanisms not as peripheral additions but as a unifying axis driving the ongoing evolution of UNet architectures, spanning architectural design, scale generalization, modality adaptation, model compression, and responsible AI integration, and establishing attention as a critical enabler of next-generation medical image segmentation systems.

9.2. Benchmarking

With its encoder-decoder, the UNet, first released in 2015, completely changed the architecture for segmenting medical images. Numerous variations, including UNet+, UNet++, UNet+++, ZNet, TNet, and WNet, have since been created [114–116]. To improve segmentation and classification tasks, these variations alter the encoder, decoder, skip connections, bridge network, and termination layer. (2) Advancements and Reviews: Reviews from 2020 to 2022 have highlighted the evolution of UNet and its applications in various medical imaging tasks [117–126]. Challenges addressed include the black-box nature of AI models, the hindrance of model size and complexity to deployment on edge devices, and the susceptibility of AI models to biases in data and model training. (3) Focus Areas of Recent Reviews: The 2021 Reviews [118,120–122], which lacked broad applicability, concentrated on particular applications such as breast tumor cell nuclei segmentation, radiation therapy planning, and orthopedic tumor identification. The 2022 Reviews [123–126] covered a range of topics, including modalities, transformer-based UNets, hybrid UNets with GANs

Table 5
Comparison table. IOA: Impact of Attention.

SN	Author	Application	# Ref.	PRISMA	IOA	Bias
1	Liu et al. [122]	Medical Image	157	x	x	x
2	Samarasinghe et al. [125]	Thorax, Neck, Abdomen	92	✓	x	x
3	He et al. [127]	Thyroid tissue	56	x	x	x
4	Wu et al. [128]	Microscopic Images	155	✓	x	x
5	Du et al. [124]	Medical Image	63	✓	x	x
6	Lagree et al. [126]	Breast Cancer	89	x	x	x
7	Punn et al. [129]	COVID-19	195	x	x	x
8	Suri et al. [71]	Vascular/non-vascular	230	✓	x	✓
9	Siddique et al. [123]	Brain tumor, Cancer, Retinal	387	x	x	x
10	Yin et al. [130]	COVID-19, Retina, Vessel	80	x	x	x
11	Saadi et al. [131]	Orthopaedic	107	x	x	x
12	Proposed	Medical/non-medical	150	✓	✓	✓

(Seg-cGAN), and methods for microscopic image analysis, as shown in Table 5. (4) Unique Contributions: Attention Mechanisms: Classifying and studying the impact of attention mechanisms (e.g., spatial-channel attention gates) on different components of UNet (encoder, decoder, skip connections, bottleneck) to enhance segmentation accuracy. (5) Addressing AI Challenges: Highlighting issues like model interpretability, deployment feasibility, and bias mitigation in AI models for medical imaging.

The evolution of UNet and its variants continues to drive innovations in medical image segmentation, with a focus on improving accuracy and generalization and addressing practical deployment challenges. Advances in attention mechanisms and hybrid architectures show promise in overcoming the limitations of traditional UNet models, making significant strides toward more effective and accessible medical imaging solutions.

9.3. A special note on computational overhead

While attention mechanisms substantially improve feature representation and segmentation accuracy in UNet-based architectures, their integration also introduces additional computational operations within the network. In particular, attention modules embedded in decoder pathways may include extra convolutional layers, pooling operations, or lightweight multilayer perceptrons that slightly increase the model's computational complexity. However, across the surveyed studies, inference time and runtime latency are not uniformly reported. When available, they are often measured under different hardware configurations, input resolutions, batch sizes, and implementation frameworks. Consequently, direct numerical comparison of inference overhead across studies is not methodologically standardized. Nevertheless, most convolution-based attention modules (e.g., spatial attention, channel attention, and serial/parallel fusion variants) are designed to be lightweight and typically add only marginal computational cost relative to the full encoder-decoder backbone. In contrast, transformer-based or global self-attention mechanisms may introduce higher memory and latency overhead, particularly in high-resolution medical imaging tasks. Therefore, the computational trade-off associated with attention integration should be interpreted within the broader context of architectural design, dataset scale, and deployment requirements.

9.4. Strengths, weaknesses, and future extensions

The proposed study provides a comprehensive narrative review of attention mechanisms in UNet-based architectures for medical image segmentation, bridging the gap between deep learning advancements and real-world clinical applications. By classifying attention mechanisms into convolution-based (cSAM, cCAM, sfSCAM, pfSCAM) and convolution-free (transformer-based) methods, we offer a structured taxonomy that enhances understanding of how attention influences various arms of UNet (encoder, decoder, skip connections, and bottleneck). The rigorous PRISMA-based study selection and statistical distribution analysis validate the broad impact of attention modules, demonstrating their effectiveness across multiple imaging modalities, including CT, MRI, ultrasound, and dermoscopy. Furthermore, we address bias and explainability (XAI), critical factors for real-world adoption, ensuring that AI-driven segmentation models are both interpretable and fair.

Despite these strengths, the study has certain limitations that future research should address. First, while we classify attention mechanisms based on their architecture and function, a deeper quantitative benchmarking (e.g., computational complexity vs. accuracy trade-off) is required to determine their optimal use cases. Second, although attention mechanisms enhance feature selection, their integration into UNet models often introduces additional computational overhead, potentially hindering real-time medical applications, especially in resource-constrained environments. Another limitation is the lack of

standardized evaluation studies, which makes it challenging to perform a unified comparison of results. Finally, including bias assessment methods is a novel contribution, but further empirical validation is needed to determine how these biases impact model generalizability across diverse datasets.

To extend this research, we propose several future directions. One key extension is the development of lightweight attention-enhanced UNet variants [131], optimized for real-time medical imaging by employing pruning [129,130], quantization [127], or knowledge distillation [128]. Another promising direction is the fusion of attention-based UNets with transformer models [132,133], leveraging global self-attention while maintaining spatial localization efficiency. Additionally, future work should focus on multi-modal learning, integrating multiple imaging modalities (e.g., MRI + histopathology) to improve segmentation accuracy and robustness. Finally, we advocate standardized benchmarking protocols that incorporate both performance (Dice, IoU, Hausdorff distance) and computational efficiency metrics to provide clearer comparisons of attention-based UNet models in clinical settings. By addressing these extensions, we can further refine the role of attention mechanisms in next-generation medical AI systems.

10. Conclusion

The attention mechanism significantly impacts various components of the UNet architecture, including the encoder, decoder, skip connections, and bottleneck. The attention mechanism in the UNet encoder helps capture more relevant and discriminative features from input images. By focusing on informative regions, the encoder can better learn the semantic context of the input, leading to more precise feature representations. The attention mechanism helps refine the decoder's feature reconstruction by selectively emphasizing the important features learned by the encoder. It enables the decoder to attend to relevant spatial regions, improving localization accuracy for segmentation tasks. Attention mechanisms facilitate better information flow through skip connections by highlighting salient features from both the encoder and decoder paths. By attending to crucial features, skip connections preserve fine details during, contributing to more accurate segmentation maps. Attention mechanisms in the bottleneck of UNet enable the model to focus on essential features while discarding noise or irrelevant information. They help compress and summarize features effectively, leading to more efficient representation learning and reduced computational overhead. Overall, integrating attention mechanisms within the

UNet architecture enhances its ability to capture meaningful features, refine feature representations, preserve spatial details, and significantly improve segmentation performance across various applications.

Ethics statement

This manuscript is a review and methodological analysis and does not involve any experiments on human or animal subjects, nor does it include the collection or use of participant data. Therefore, ethical approval, reference numbers, approval dates, and informed consent are not applicable to this study. All works discussed were conducted in compliance with relevant laws and institutional guidelines as reported in the original publications, and no identifiable personal or clinical information is used in this manuscript.

Funding

The research received no external funding.

CRediT authorship contribution statement

Siddharth Gupta: Formal analysis, Methodology, Software. **Yogendra Chhetri:** Conceptualization, Investigation, Writing – original draft. **Amit Pandey:** Data curation, Visualization, Writing – original draft, Writing – review & editing. **Akansha Singh:** Formal analysis, Project administration, Writing – review & editing. **Mrinalini Bhagwati:** Software, Validation, Visualization. **Suneet K. Gupta:** Data curation, Resources, Validation. **Neha Gupta:** Formal analysis, Methodology, Project administration. **Surbhi Verma:** Data curation, Formal analysis. **Ekta Tiwari:** Investigation, Resources, Writing – review & editing. **Mustafa Al-Maini:** Investigation, Supervision, Validation. **Mostafa M. Fouda:** Formal analysis, Supervision, Validation. **Gavino Faa:** Formal analysis, Investigation. **Esma R. Isenovic:** Supervision, Validation. **Andrew Laine:** Conceptualization, Investigation, Methodology. **Luca Saba:** Methodology, Project administration, Validation. **Jasjit S. Suri:** Conceptualization, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

Convolution Free Attention

Vaswani et al. [30] proposed a new architecture transformer, as shown in Figures A1 and A2, which is purely based on the self-attention mechanism. Transformers eliminate convolution and recurrence and are now also used with UNet for image segmentation. The main components of the Transformer are the encoder, decoder, multi-head self-attention, and encoder–decoder attention modules. Each encoder includes a self-attention layer and a feed-forward network, with the output from one encoder passed to the next. Key (K) and value (V) vectors are generated from the encoder output and used by the decoder to focus on relevant input features. The final prediction is produced by applying a linear layer and a SoftMax function to the decoder's output. The self-attention mechanism's output is determined by applying Equation a1.

$$SA = \psi \left(QxK^T / \sqrt{d_k} \right) \quad (a1)$$

where Q is the query vector, K is the key vector, and ψ is the SoftMax activation function. In this mechanism, each input generates query (Q), key (K), and value (V) vectors through learned linear projections. Each vector typically has dimension d_k (for instance, 64), learned during training. Attention scores are computed as scaled dot products between Q and K, normalized by SoftMax to produce weights that sum to one. These weights are then applied to the value vectors, yielding a context-aware representation that captures global dependencies within the input sequence.

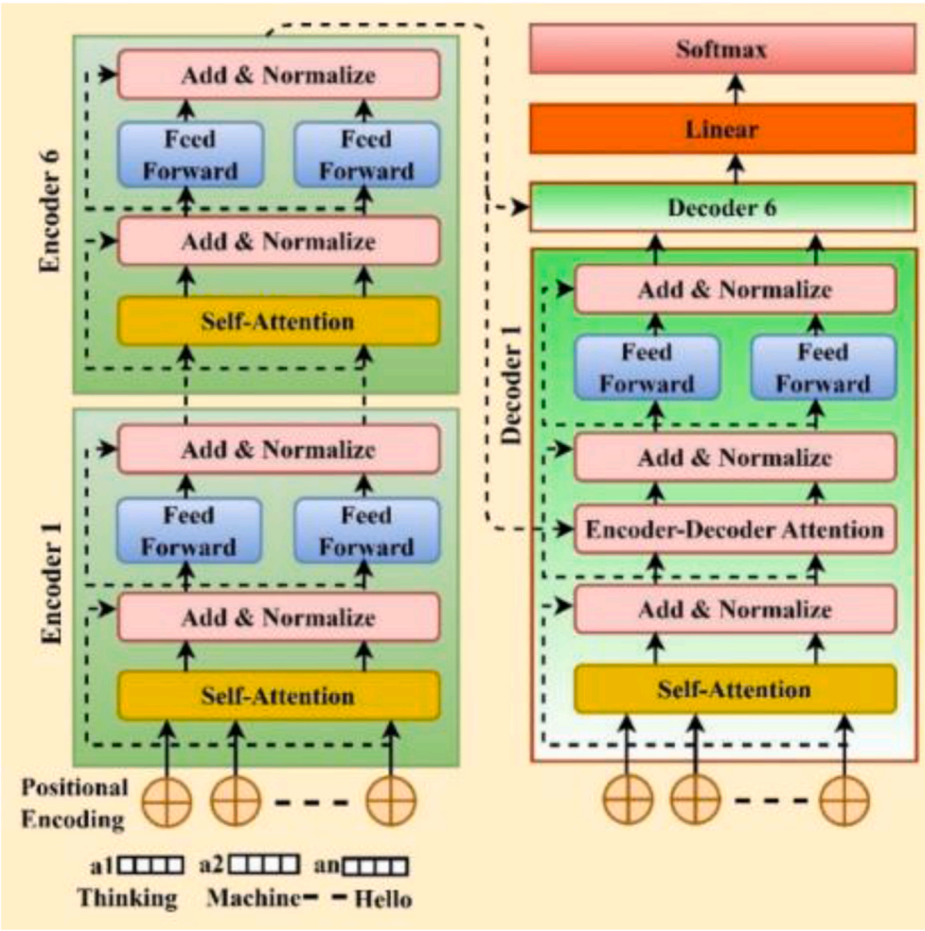


Fig. A1. Transformer architecture in detail.

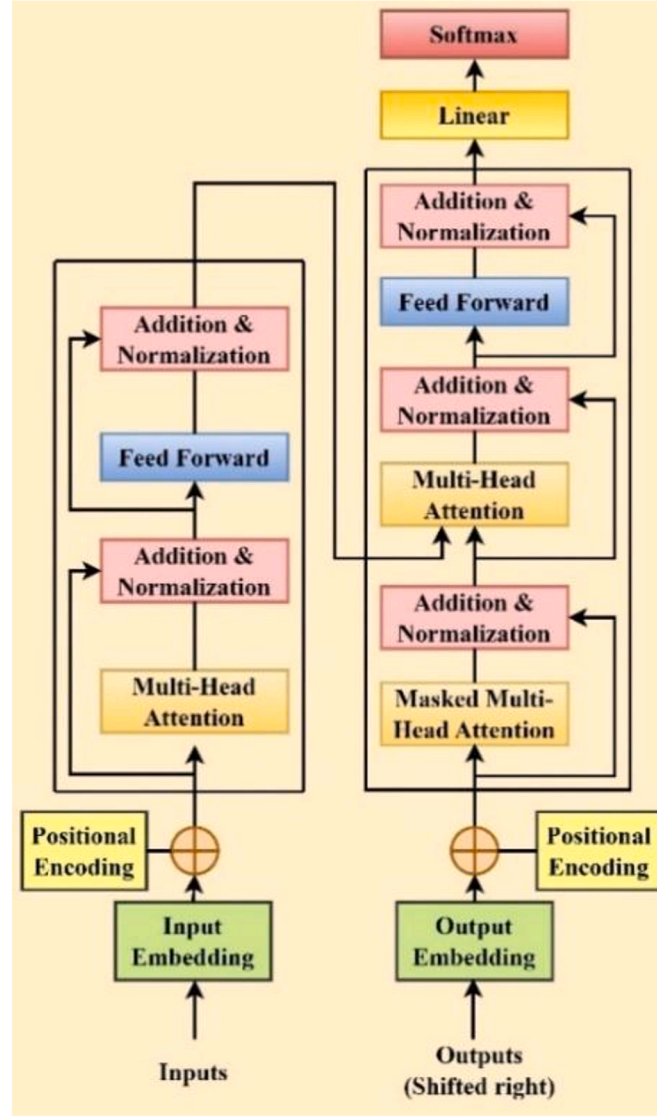


Fig. A2. Abstract view of transformer architecture.

Appendix B

Impact of attention mechanism in the multi-armed of UNet

In this section, we discuss the impact of cSAM, cCAM, sfSCAM, and pfSCAM on the multi-arm architecture of UNet. Incorporating these attention mechanisms significantly enhances segmentation performance by refining feature extraction. cSAM and cCAM prioritize channel-wise features, while sfSCAM and pfSCAM focus on spatial and positional information for complex medical imaging tasks. At the end of this section, we explore applications that benefit from these attention mechanisms in the encoder of the UNet (Table B1), covering various image modalities such as MRI, CT, and ultrasound. Additionally, we discuss the loss functions used, including Dice loss, cross-entropy loss, and focal loss, essential for optimizing model performance.

B.1 Impact of convolution-based Spatial Attention Module on the multi-armed UNet

cSAM is used to extract the relevant parts of an image while ignoring the irrelevant parts. Gu et al. [134] proposed a novel multi-attention cSAM called joint-cSAM, inspired by non-local networks [135] and attention-gated networks [39]. In this paper, the authors used four cSAM blocks named (SAD1-4) to extract attention maps at four different resolution levels, as described in Figure B1. SAD₁ is used at the lowest resolution level to capture interactions between all pixels, uses the concept of non-local block, and captures interactions between all pixels using the concept of non-local blocks.

It also uses the concept of a non-local block and the detailed structure of SAD_1 described in Figure B2. Suppose that $\hat{x}$ represents the FM with dimension $256 \times H \times W$, where H is height, W is width, and 256 is the channel number. To reduce the dimension of FM, $\hat{x}$, first apply three parallel 1×1 convolutions with 64 channels and obtain three reduced FMs named $\hat{x}', \hat{x}'', \hat{x}'''$ with dimension $64 \times H \times W$. In the next step, these three FMs are reshaped with dimension $64 \times HW$. The cSAM coefficient was computed using the following formula, in which $\alpha_1 \in (0, 1)^{HW \times HW}$ was a square matrix, and T was a transpose operation. To guarantee that all the pixels interact, each pixel's feature was represented by α_1 , a weighted sum of all the pixels' features. To ensure that a portion of every row is equal to one, σ was a row-wise softmax function.

$$\alpha = \sigma(\hat{x}'^T \cdot \hat{x}'') \quad (54)$$

The calibrated FM is then computed using the equation below. After reshaping $\hat{a}$ to $64 \times H \times W$, the authors expanded $\hat{a}$ to match the channel number of a used Φ^{256} , a 1×1 convolution with BN, and an output channel number of 256. In the end, a residual link is employed to speed up the spread of information during training, and the SAD_1 output is as follows:

$$\hat{x} = \alpha_1 \cdot \hat{x}'''^T \quad (55)$$

$$y_{SAD_1} = \Phi^{256}(\hat{x}) + x \quad (56)$$

Due to the increased memory consumption, the non-local block was applied to FM with high resolution, and AG [39] was used to learn the attention coefficient in SAD_{2-4} . Single-pathway AG (Figure B3) leads to a noisy attention map, so the authors proposed dual-pathway attention, which helped remove the noise and improved segmentation robustness, as shown in Figure B3. Let x^l represent the low-level FM at the encoder side at scale s , and x^h represent the high-level FM at the decoder side, and the level was $s+1$ with lower spatial resolution. The 1×1 convolution with channel number C was used to compress x^l and x^h , as shown in Figure B4. The outputs were then summed, and the ReLU activation function was applied. To obtain the pixel-wise attention coefficient $\alpha \in (0, 1)^{H \times W}$, the outputs of the ReLU steps were fed back into a 1×1 convolution with one output channel and then processed using a sigmoid process. Following this stage, the low-level FM x^l is multiplied by α to obtain its calibration. The attention map in dual-pathway AG is depicted as follows. The output of four dual-pathway AG for SAD_{2-4} is depicted in Figure B4 and is explained as follows:

$$y_{SAs} = \text{ReLU}[\Phi^C((x^l \cdot \hat{a}) \oplus (x^h \cdot \hat{a}))] \quad (57)$$

where $\oplus$ represents channel concatenation, Φ^C denotes 1×1 convolution with BN, and C is the number of channels.

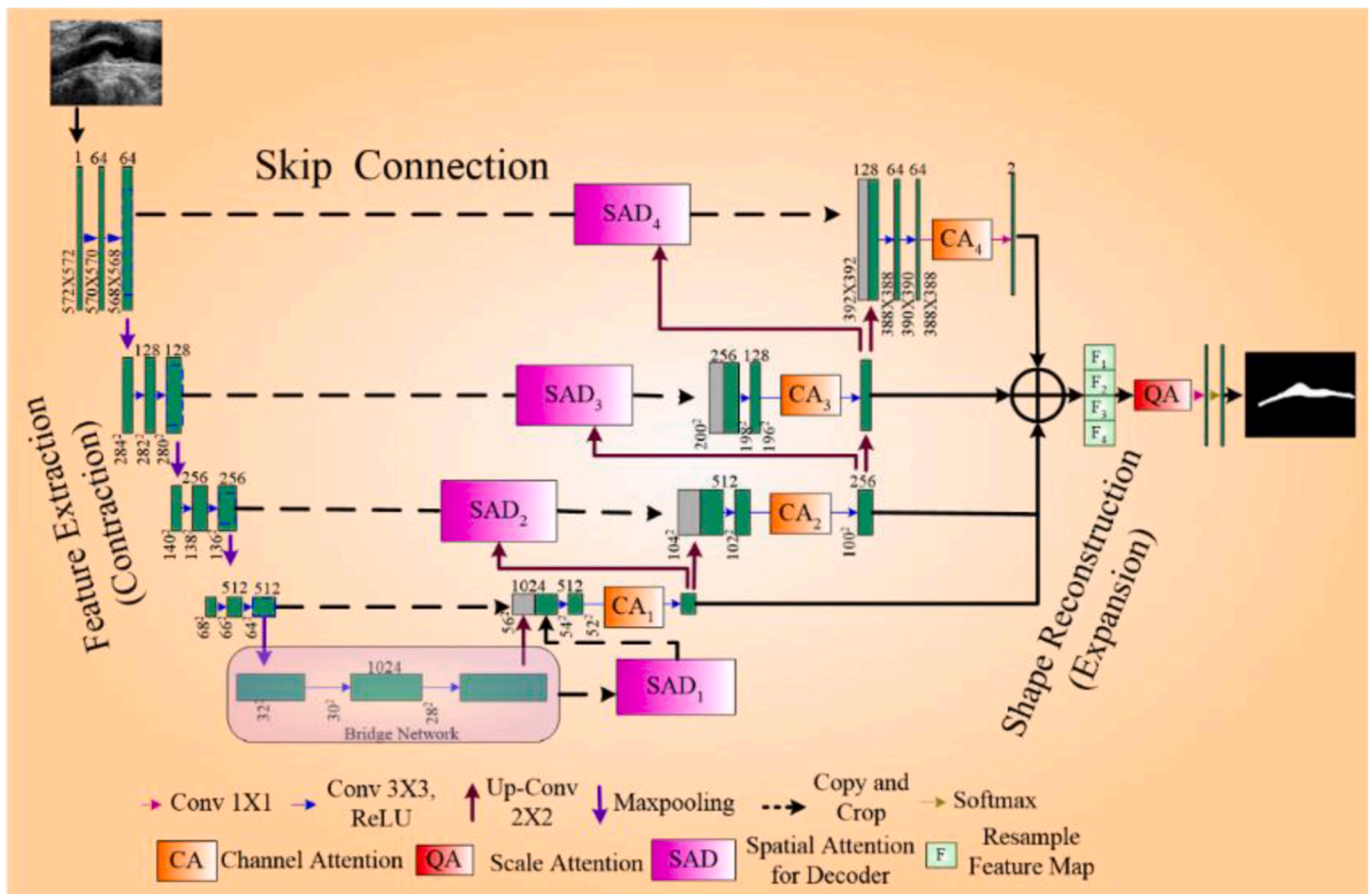


Fig. B1. Architecture of comprehensive attention CNN (CA-Net).

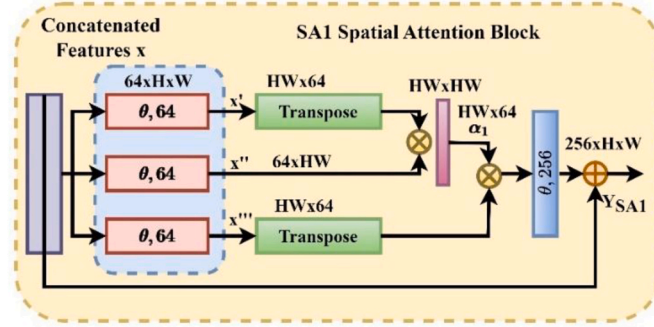
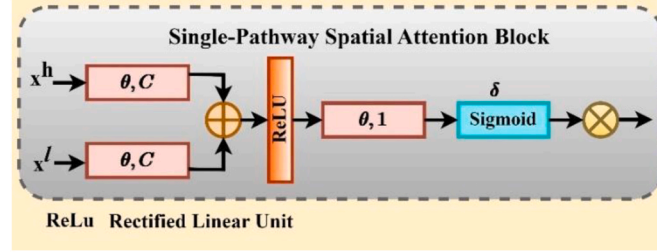
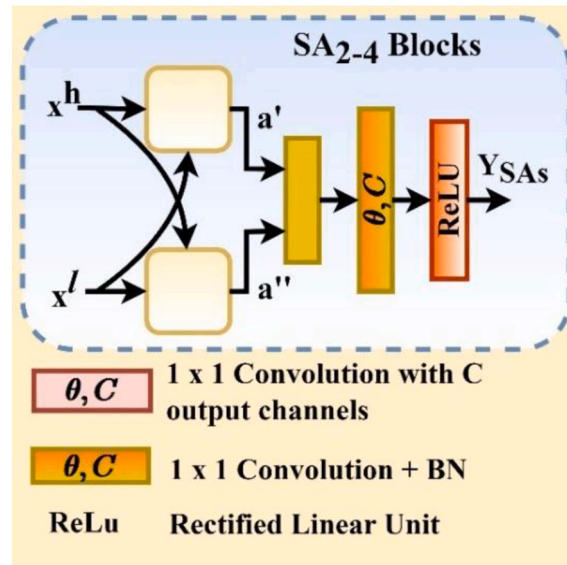
Fig. B2. Spatial attention block (SAD₁).

Fig. B3. Single-pathway spatial attention (SAD).

Fig. B4. Dual-pathway spatial attention block (SAD₂₋₄).

B.2 Impact of Convolution-Based Channel Attention Module on the multiarm of UNet

Because of the complex facial picture, uneven tongue texture, lips, and face, traditional methods could not segment the tongue accurately. To tackle these issues, Song et al. [136] presented RAFF-Net, an autonomous tongue region segmentation network based on residual attention networks and multiscale feature fusion. Utilizing the SE [24] block with varying numbers of ResBlock served as an encoder to extract layered features from images based on the UNet backbone network. Both the number of parameters in the network model and the decoder structure of UNet were reduced. In the meantime, to further increase the detection accuracy, the multiscale feature fusion module was created to optimize the network parameters by adding a bespoke loss function instead of the standard cross-entropy loss function. By introducing an attention mechanism, neural networks can limit their attention to relevant information, focus on the information most important to the current task amid a large amount of input, and perform tasks more accurately and efficiently.

The SE block can highlight significant characteristics, suppress superfluous features, and learn the link between the model's channels. To apply the cCAM, SE explicitly models the weights, weighting the original FM to give each channel a distinct level of priority, and scales the original FM to give each channel a distinct level of priority. The FM dimension of the current feature layer is $W \times H \times C$. Image features were retrieved from a single image as a starting point. The size of the pooled FM changes from $W \times H \times C$ to $1 \times 1 \times C$. The $[H, W]$ dimension of the FM is AP or MP (AP was used in this work). The action proceeds through the ReLU function after passing through a fully connected layer, which reduces the number of channels from C to C/r . The hyperparameter "r" controls the SE block's computational load and has varying effects on network performance. The experimental section will address the value of r in detail. After that, it passes through a fully linked layer where the number of channels is increased from C/r to C . A sigmoidal function

then appears right after. Following the acquisition of a $1 \times 1 \times C$ weight for every channel C , the weights are applied to the weight of each channel multiplied by itself.

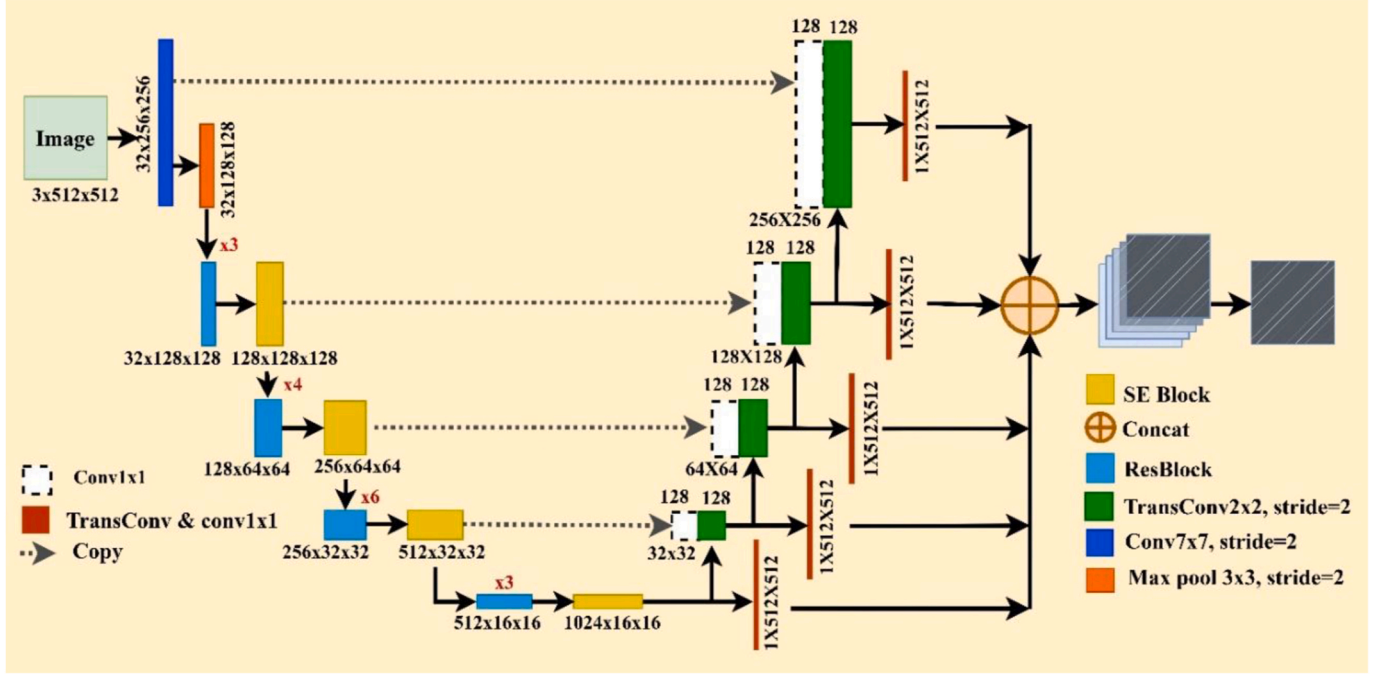


Fig. B5. Impact of cCAM on mUNet.

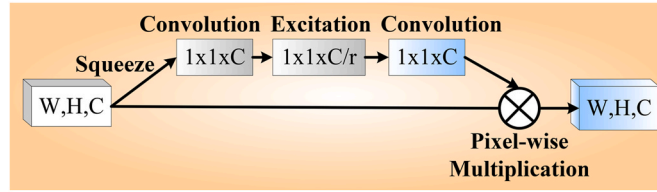


Fig. B6. Squeeze and excitation network.

B.3 Impact of sfSCAM on the multiarm of UNet

To achieve effective and consistent segmentation of aneurysm images obtained via MRA, Yuan et al. [137] introduced DCAUNet despite considerable variations in size. This image segmentation model leverages a network with dense convolution and an attention mechanism, integrating rich, detailed semantic information. The foundational architecture is a modified UNet model that combines the cCBAM and densely connected blocks. The dense block consists of a convolution layer, a ReLu activation, and a BN layer. Its objectives include improving the network's training efficiency, alleviating vanishing gradients, and capturing complex aneurysm features, as shown in Figure B7. Multi-scale fusion is carried out during the up-sampling process. This approach addresses the wide size variability of aneurysms by leveraging the SCAM components of cCBAM to enhance feature discrimination and improve the acquisition of essential information, as shown in Figure B8.

As shown in Figure B9, the channel attention sub-module examines the relationships between features. To generate the spatial context descriptors G_{AP}^C and G_{MP}^C . The FM and FC undergo GAP and GMP. This changes the feature dimensions from $C \times D \times H \times W$ to $C \times 1 \times 1 \times 1$. These descriptors are fed into a shared cMLP. The concatenated output features are combined using sigmoid activation to form the spatial FM, N^C . The dimensions of the FM are $C \times 1 \times 1 \times 1$. For N^C , the mathematical formulation was:

$$N^C(F) = \Theta(\text{cMLP}(\text{AP}(G)) + \text{cMLP}(\text{MP}(G))) = \Theta(X_1(X_0(G_{AP}^C)) + X(X_0(G_{MP}^C))) \quad (58)$$

where the sigmoidal function is denoted by Θ . It should be noted that X_0 comes after the ReLu activation function, and the cMLP weights, X_0 and X_1 , are shared for both inputs. The first input FM, G^C is multiplied by N^C to get G^C the CAM's output. Figure B10 depicts the intricate arrangement of the spatial attention submodules. Using the FM, G^C from the CAM output as GMPS maps: GMPS, input by executing GAP and GMP. The FM's dimensions shift from $C \times D \times H \times W$ to $1 \times D \times H \times W$. Concatenating G_{AP}^S and G_{MP}^S yielded S , whose dimensions are $2 \times D \times H \times W$. The dimensionality was then reduced to a channel number for getting S' , or $1 \times D \times H \times W$, via the $7 \times 7 \times 7$ convolution. The FM has the following dimensions: $1 \times D \times H \times W$. S' derives N^S from the sigmoid activation function. It measured $1 \times D \times H \times W$. It is defined as:

$$N^S(G) = \Theta(g^{7 \times 7 \times 7}([\text{AP}(G); \text{MP}(G)])) = \Theta(g^{7 \times 7 \times 7}([G_{AP}^S; G_{MP}^S])) \quad (59)$$

where G_{AP}^S denotes the AP feature over the channel, G_{MP}^S denotes the largest pooling characteristic from the channel, and $g^{7 \times 7 \times 7}$ demonstrates the

convolution process utilizing a $7 \times 7 \times 7$ convolutional filter.

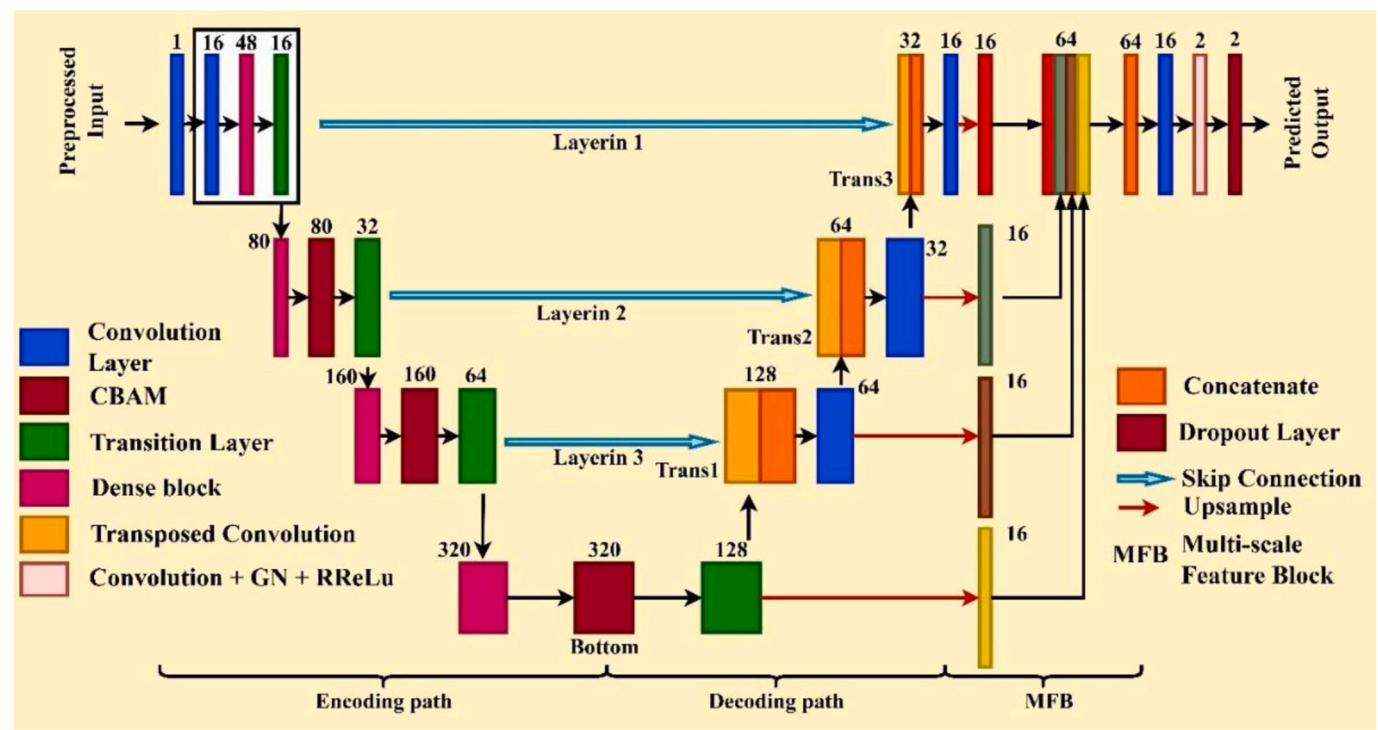


Fig. B7. Impact of sfSCAM on mUNet.

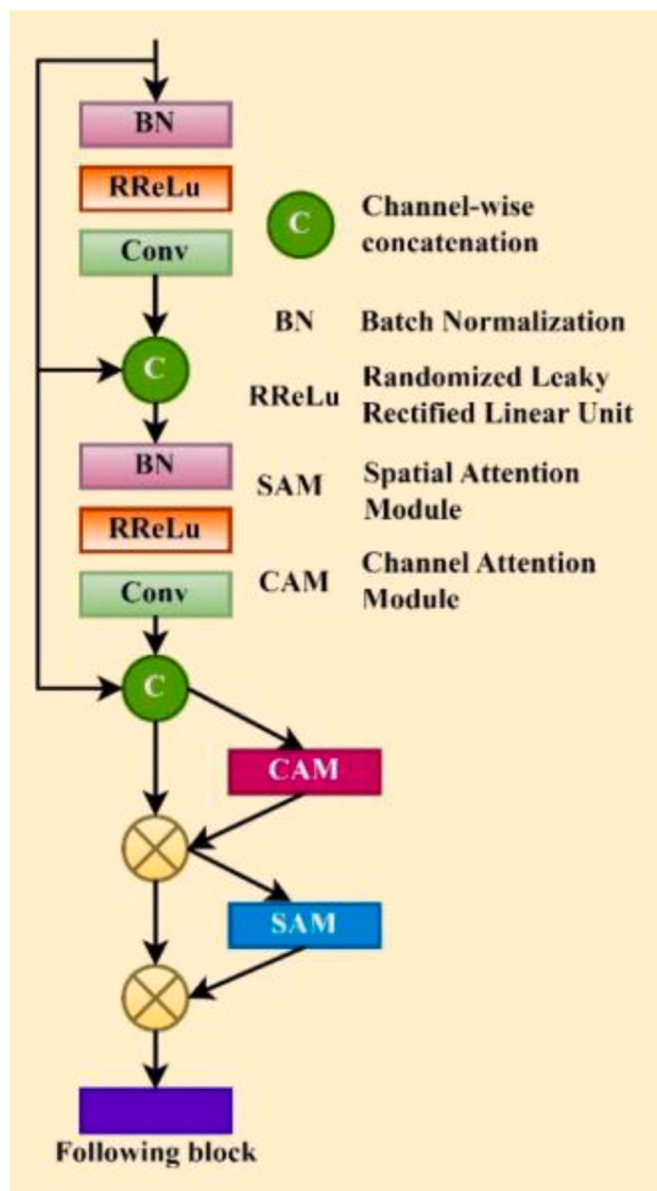


Fig. B8. Incorporating dense block and cCBAM.

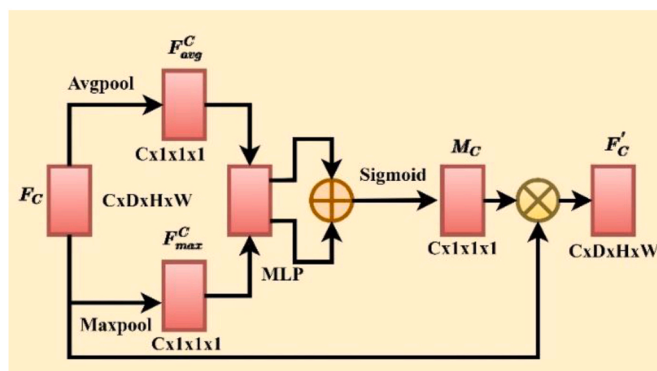


Fig. B9. Channel attention module.

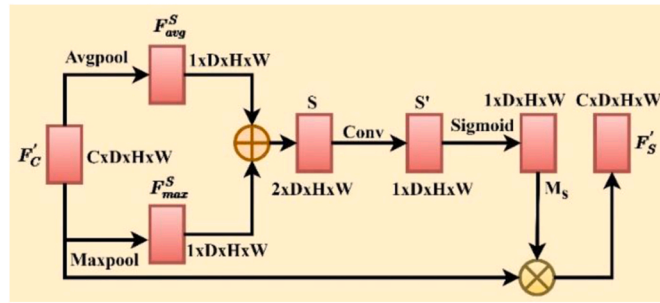


Fig. B10. Spatial attention module.

B.4 Impact of pfSCAM on the multiarm of UNet

To segregate brain cancers using Magnetic Resonance image (MRI) and forecast the survival rate using a variety of machine learning techniques, Islam et al. [138] created an attention CNN. The authors performed segmentation using a 3D UNet architecture by integrating channel and spatial attention simultaneously. Spatial and channel attention have always helped. Therefore, to construct 3D spatial and channel attention, we employ 3D inter-spatial and inter-channel feature interaction and introduce 3D attention units, as shown in Figure B11. To acquire the 3D attention map, we first perform a $1 \times 1 \times C$ convolution into the $H \times W \times 1$ dimension. We perform AP in parallel and feed the outcomes into the neural network to determine the $1 \times 1 \times C$ channel correlation. The 3D attention map encodes rich spatial and channel attention. Additionally, the authors employed a fused skip connection to reduce the singularity and sparsity induced by these parallel excitations. Adding skip connections improves segmentation prediction and makes learning more general. Finally, the impact of pfSCAM on mUNet is illustrated in Figure B12, highlighting the improvements brought by these attention mechanisms.

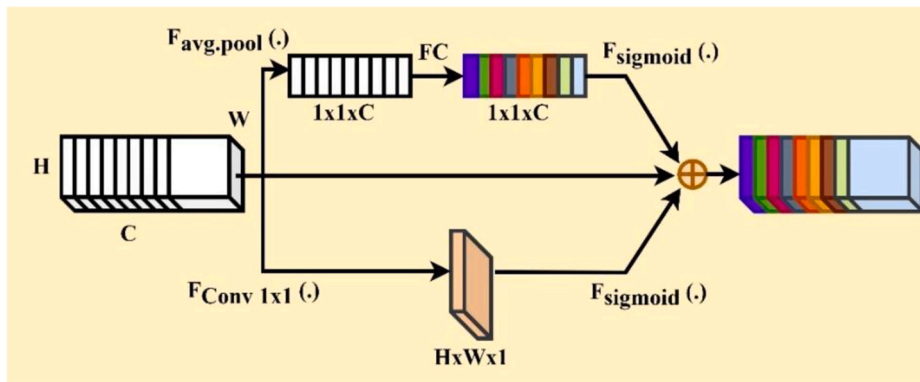


Fig. B11. 3D spatial and channel attention with skip connection.

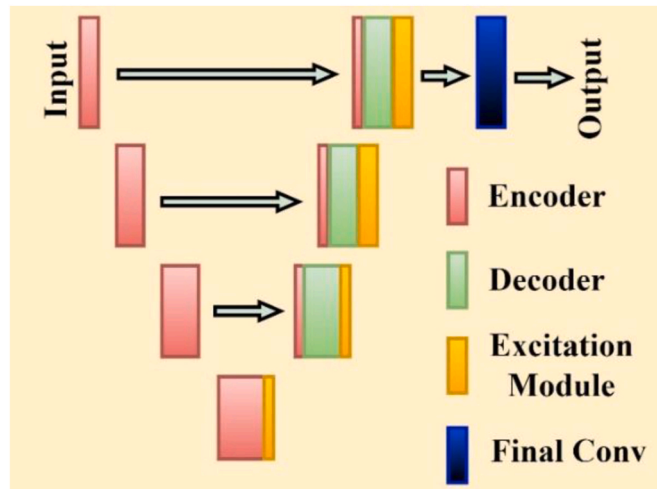


Fig. B12. Impact of pfSCAM on mUNet.

Table B1

Segmentation of medical images with the help of an attention mechanism on bUNet.

SN	Author		FOV (Application)		LF Type
M1	Gu et al. [134]	MRI/Dermoscopy	Skin lesion/Fetal Brain MRI	cSAM, Ccam	Dice loss
M2	Song et al. [136]	Satellite/Camera	Toungue segmentation	Ccam	Dice/CE/MSD
M3	Amer et al. [139]	CT	Lung and left ventricle	cSAM, cCAM	Dice Loss
M4	Guo et al. [137]	Satellite/Camera	RetinAal vessel	Ccam	BCE
M5	Gandhi et al. [140]	MRI	Brain MRI	PFSCAM	Dice loss
M6	Xu et al. [141]	Satellite/Camera	Nuclei Segmentation	cCAM	dice loss/Focal loss
M7	Xu et al. [142]	MRI	Polyp/nuclei/lesion/Brain/Myeloma	cCAM	Dice
M8	Yuan et al. [143]		Aneurysm segmentation	sfSCAM	Dice
M9	Jing et al. [144]	Satellite/Camera	Road Crack	sfSCAM	Dice
M10	Bourigault et al. [145]	CT	Head and neck tumor	cCAM	Dice/focal loss
M11	Chang et al. [146]	Satellite/Camera	Impervious	sfSCAM	Dice
M12	Islam et al. [138]	MRI	Brain tumor	pfSCAM	Dice
M13	Lee et al. [147]	Satellite/Camera	change detection in satellite	pfSCAM	CE
M14	Li et al. [148]	Satellite/Camera	Remote sensing	cCAM	CE

Appendix C

One problem has many solutions

Appendix C illustrates that a single medical segmentation problem may admit multiple structurally valid attention-based solutions. Rather than prescribing a universally optimal mechanism, the selection of an attention fusion strategy should be guided by modality-specific characteristics such as target size, boundary contrast, class imbalance, and contextual complexity. Spatial attention mechanisms are often suited for fine boundary localization, channel attention for feature recalibration in heterogeneous tissues, and parallel fusion strategies for multi-scale contextual modeling. Therefore, attention modules serve as structural design principles that can be tailored to the clinical and imaging characteristics of the target application.

To concretely demonstrate this principle, we present disease-wise examples showing that multiple attention-based UNet variants can effectively address the same clinical task. For example, if a particular disease is lung cancer, then it can be solved by pfSCAM [28], sfSCAM [89], cSAM [76], and cCAM [139]. In the case of brain cancer, the solutions are pfSCAM [48], sfSCAM [72], cSAM [80], and cCAM [134]. If the problem is liver cancer, then the solutions are cSAM [63], pfSCAM [26]. The other applications are skin lesions, kidney, eye, pancreas, and cell nuclei, and their solutions are described in Tables C1–C4.

Table C1

One Problem with many solutions using attention-UNet models. S: Skip; E: Encoder; D: Decoder; B: Bottleneck; M: Multi-arm; AOE:Arm of UNet.

Appli- cation	Model	Author	AOE	SEN	ACC (%)	PREC (%)	F1-Score (%)	DICE Coefficient (%)	mIOU
Lung	PFSCAM	Hou et al. [26]	E	.873/.872/.966	.996/.989/.978	.995/.976/N	.930/.921/N	.874/.926/.963	.805/.863.940
	SFSCAM	Khanh et al. [72]	S	N	N	72.43+-7.14/ 51.79+-4.42/ 83.74+-1.81	N	73.31+-5.85/56.26+- 3.51/85.83+-1.41	N
	CSAM	Cheng et al. [73]	S	N	N	N	N	.8711+-0.1114	N
	PFSCAM	Fang et al. [92]	D	.9823/.9902/.817/.809	.9946/.9932/.948/.948	N	.9841/.9800/.870/.869	N	N
	SFSCAM	Xu et al. [107]	D	.9010/.6894/.9660	.9604/.9416/.9872	.8938/.6862/.9806	.8970/.6858/.9730 [149].9730	.8848/.7132/.9722	N
	CSAM, CCAM	Amer et al. [139]	M	N	N	N	98.9+-5.3/ 97.5+-4.3	98.3+-2.4/96.7+-3.5	N
	PFSCAM	Hou et al. [26]	E	.873/.872/.966	.996/.989/.978	.995/.976/N	.930/.921/N	.874/.926/.963	.805/.863.940
Liver	CSAM	Li et al. [63]	S	N	N	.98	N	98.15%	97.48%
	CSAM	Li et al. [65]	S		N	.9185/.9519/.9400/.9423	N	98.15%/94.31%/ 94.79%/93.55%	97.48/89.23/90.10/ 87.89
	PFSCAM	Liu et al. [69]	S	N	N	N	N	.9723/.8816/.9707/ .8952/.8452/.8696/ .9554/.9083	.9506/.8127/.9473/ .8340/.7506/.7857/ .9214/.8628
	CCAM	Wang et al. [79]	S	.9743	.9934/.9745	.8341	N	.9797/.9081	N
	PFSCAM	Hong et al. [111]	B	N	N	N	N	96.13+-1.03/85.90+- 5.63	N

Table C2

One Problem with many solutions using attention-UNet models. S: Skip; E: Encoder; D: Decoder; B: Bottleneck; M:Multi-arm; AOE:Arm of UNet.

Appli- cation	Model	Author	AOE	SEN	ACC (%)	PREC (%)	F1-Score (%)	DICE Coefficient (%)	mIOU
Brain	PFSCAM, SFSCAM	Hatamizadeh et al. [48]	E	N	N	N	N	.822+-176	N
	SFSCAM	Khanh et al. [72]	S	Are u get this onlN	N	72.43+-7.14/ 51.79+-4.42/ 83.74+-1.81	N	73.31+-5.85/ 56.26+-3.51/ 85.83+-1.41	N
	CSAM	Aboussaleh et al. [80]	S	N	99.1	N	N	87.41/80.61/70.33	77.64/67.63/ 54.24
	PFSCAM	Agarwala et al. [81]	S	.73/.90, .80	N	N	N	.73/.90/.83	N
	CCAM	Huang et al. [100]	D	N	N	N	N	90.9	N
	CCAM	Noori et al. [106]	D	N	N	N	N	.813/.895/.823	N
	CSAM, CCAM	Gu et al. [134]	M	N	N	N	N	87.72+-4.43/95.51/ +-2.66	N
	PFSCAM	Gandhi et al. [140]	M	N	N	N	N	93.46+-1.05/ 96.81+-1.15/ 94.28+-2.01	N
	CCAM	Xu et al. [141]	M	N	.990+-015/ .950+-045/ .959+-045	.917+-148/ .871+-113/ .914+-098	N	.916+-141/.886+- .078/.914+-077	.861+-156/ .806+-121/ .850+-114
	PFSCAM	Islam et al. [138]	M	N	N	N	N	70.47/89.8/79.27	N
Skin	SFSCAM	Li et al. [50]	E	.939/.9328/ .9591	N	N	N	.9461/.9523/.9172	.8892/.9615/ .9025
	CSAM	Li et al. [64]	S	.9112/ .9493/.8335	N	N	.8955/ .8831/ .6110	88.59%/85.91%/ 72.48%	82.32%/87.74%/ 66.57%
	CSAM	Abraham et al. [68]	S	N	N	.829+-027/ .858+-020	N	.804+-024/.856+- .007	N
	CSAM	Lin et al. [75]	S	89.58/ 91.10/ 91.97/93.34	99.36/98.23/ 95.76/93.21	N	N	96.02/93.86/91.72/ 87.75	N
	CSAM	Pennisi et al. [91]	D	N	N	N	N	.9035/.9301	N
	PFSCAM	Fang et al. [92]	D	.9823/ .9902/.817/ .809	.9946/.9932/ .948/.948	N	.9841/ .9800/.870/ .869	N	N
	CSAM	Liu et al. [86]	D	N	N	N	N	.950/.939/.919	.908/.893/.860
	SFSCAM	Xu et al. [107]	D	.9010/ .6894/.9660	.9604/.9416/ .9872	.8938/.6862/ .9806	.8970/ .6858/ .9730	.8848/.7132/.9722	N
	CSAM, CCAM	Gu et al. [134]	M	N	N	N	N	87.72+-4.43/95.51/ +-2.66	N

Table C3

One Problem with many solutions using attention-UNet models. S: Skip; E: Encoder; D: Decoder; B: Bottleneck; M:Multi-arm; AOE: Arm of UNet.

Appli- cation	Model	Author	AOE	SEN	ACC (%)	PREC (%)	F1-Score (%)	DICE Coefficient (%)	mIOU
Eye	PFSCAM	Wang et al. [53]	E	N	N	N	N	N	85.42
	CSAM	Wu et al. [66]	S	.7798	.9615	N	N	N	N
	CSAM	Guo et al. [110]	B	.8212/ .8573	.9698/.9775	N	.8263/.8153	N	N
	CCAM	Guo et al. [137]	M	.8058/ .8435	.9691/.9744	N	.8202/.8063	N	N
Gastic Cancer	CSAM	Oktay et al. [23]	S	N	N	.848+-110	N	.840+-087	N
	CSAM	Schlemper et al. [61]	S	N	N	.849+-098	N	.840+-087	N
Pancreas	CSAM	Oktay et al. [23]	S	N	N	.848+-110	N	.840+-087	N
	CSAM	Schlemper et al. [61]	S	N	N	.849+-098	N	.840+-087	N
	PFSCAM	Liu et al. [69]	S	N	N	N	N	.9723/.8816/.9707/ .8952/.8452/.8696/ .9554/.9083	.9506/.8127/.9473/ .8340/.7506/.7857/ .9214/.8628
Cell Nuclei	CSAM	Li et al. [64]	S	.9112/ .9493/ .8335	N	N	.8955/ .8831/.6110	88.59%/85.91%/ 72.48%	82.32%/87.74%/ 66.57%

(continued on next page)

Table C3 (continued)

Appli- cation	Model	Author	AOE	SEN	ACC (%)	PREC (%)	F1-Score (%)	DICE Coefficient (%)	mIOU
Kidney	CSAM	Zhang et al. [74]	S	N	N	.94+-06	N	.93+-04	N
	CCAM	Zeng et al. [99]	D	N	N	N	.8278	.8008	N
	CCAM	Xu et al. [141]	M	N	.990+-015/ .950+-045/ .959+-045	.917+-148/ .871+-113/ .914+-098	N	.916+-141/.886+- .078/.914+-077	.861+-156/.806+- .121/.850+-114
	CSAM	Li et al. [65]	S		N	.9185/.9519/ .9400/.9423	N	98.15/94.31/94.79/ 93.55	97.48/89.23/90.10/ 87.89
	PFSCAM	Liu et al. [69]	S	N	N	N	N	.9723/.8816/.9707/ .8952/.8452/.8696/ .9554/.9083	.9506/.8127/.9473/ .8340/.7506/.7857/ .9214/.8628
Breast	CSAM	Tang et al. [67]	S	N	97.33+-14/ 99.27+-01	84.81+-1.32/ 92.02+-13	84.16+-47/ 91.71+-17	N	73.27+-43/84.75+- .30
	CSAM	Abraham et al. [68]	S	N	N	.829+-027/ .858+-020	N	.804+-024/.856+-007	N
	SFSCAM	Xu et al. [107]	D	.9010/ .6894/ .9660	.9604/.9416/ .9872	.8938/.6862/ .9806	.8970/ .6858/.9730	.8848/.7132/.9722	N
Aneurysm	CSAM	Su et al. [71]	S	N	N	N	N	88.50-90.71 %	N
	SFSCAM	Yuan et al. [143]	M	.7881/ .8276	N	.7881	N	.7455/.8412	N

Table C4

One Problem with many solutions using attention-UNet models. S: Skip; E: Encoder; D: Decoder; B: Bottleneck; M: Multi-arm; AOE:Arm of UNet

Appli- cation	Model	Author	AOE	SEN	ACC (%)	PREC (%)	F1-Score (%)	DICE Coefficient (%)	mIOU
Polyp	SFSCAM	Khanh et al. [72]	S	N	N	72.43+-7.14/ 51.79+-4.42/ 83.74+-1.81	N	73.31+-5.85/ 56.26+-3.51/ 85.83+-1.41	N
	CSAM	Lin et al. [75]	S	89.58/ 91.10/ 91.97/93.34	99.36/98.23/ 95.76/93.21	N	N	96.02/93.86/91.72/ 87.75	N
	CSAM	Chenerlogh et al. [78]	S	N	N	N	N	95.79+-01/93.84+- .03/83.00	N
	CSAM	Liu et al. [98]	D	N	N	N	N	.950/.939/.919	.908/.893/.860
	CCAM	Xu et al. [142]	M	N	.990+-015/ .950+-045/ .959+-045	.917+-148/ .871+-113/ .914+-098	N	.916+-141/.886+- .078/.914+-077	.861+-156/ .806+-121/ .850+-114
Melonoma	CSAM	Pennisi et al. [91]	D	N	N	N	N	.9035/.9301	N
Head & Neck	CSAM	Liu et al. [98]	D	N	N	N	N	.950/.939/.919	.908/.893/.860
	SFSCAM	Zhong et al. [97]	D	N	N	N	N	81.88	N
	CCAM	Bourigault et al. [145]	M	N	N	N	N	.753	N
Satellite	SFSCAM, PFSCAM	Adil et al. [101]	D	N	N	92.27	90.02	N	82.68
	PFSCAM	Zhang et al. [102]	D	N	85.11	89.92/78.10/ 75.16	88.71/ 78.81/ 78.58	N	69.82
	PFSCAM	Lee et al. [147]	M	N	N	N	.88/.83/ .59/.40	N	N
Road Crack	CCAM	Fan et al. [103]	D	N	N	N	N	N	89.1
	PFSCAM	Gao et al. [104]	D	N	N	N	66.59/ 56.49	N	49.92/39.36
	SFSCAM	Jing et al. [144]	M	N	N	88.91/63.2/58.9	87.2/ 71.1/55.4	N	N

Appendix D

Acronym Table

SN	Abb*	Definition	SN	Abb*	Definition
1	AF	Attention Fusion	20	FOV	Field of view
2	AG	Attention gate	21	HSD	Hausdorff surface distance

(continued on next page)

(continued)

SN	Abb*	Definition	SN	Abb*	Definition
3	AMFF	Attention-modulated feature fusion	22	LM	Low Moderate
4	AmUNet	Attention in the multibranch of UNet	23	MH	Moderate High
5	AsUNet	Attention in the skip of UNet	24	MLP	Multilayer perceptron
6	ASPP	Atrous spatial pyramid pooling	25	pfSCAM	Parallel fusion of spatial and channel attention module
7	bUNet	Bottleneck of UNet	26	PSNR	Peak-signal-to-noise ratio
8	BN	BatchNorm	27	QAUNet	Quartet attention UNet
9	CA	Channel Attention	28	RBA	Regional-bias area
10	CBAM	Convolution block attention module	29	ReLU	Rectified linear unit
11	cCAM	Convolution Attention Module Attention Module	30	RMSE	Reading society of model engineers
12	CNN	Convolution Neural Network	31	SA	Spatial Attention
13	CV	Cross-validation	32	SCAM	Spatial and Channel attention module
14	CVD	Cardiovascular disease	33	sSAM	Convolution-based Spatial attention module
15	dUNet	Decoder of UNet	34	sfSCAM	Serial fusion of spatial and channel attention module
16	DNN	Deep neural network	35	SSIM	Structural similarity index measure
17	eUNet	Encoder of UNet	36	VD	Venn Diagram
18	FC	Fully connected	37	XAI	Explainable AI
19	FFM	Feature Fusion Module			

Symbol Table

SN	Sym*	Definition	SN	Sym*	Definition
1	α	Weight parameter	21	SA	Self-attention mechanism
2	$f3 \times 3$	3x3 convolution filter	22	Q	Query vector
3	$f5 \times 5$	5x5 convolution filter	23	K	Key vector
4	$f1 \times 1$	1x1 convolution filter	24	$\alpha1$	The coefficient in the attention map
5	fDB	Depth-wise convolution filter	25	HW	Height and Width of the feature map
6	fBN	Batch normalization filter	26	σ	Row-wise softmax function
7	Fmap	Spatial attention map	27	a	Feature map with dimension 256xHxW
8	θ	Sigmoid activation function	28	x', x'', x'''	Reduced feature maps with dimension 64xHxW
9	Fs	Final spatial attention output	29	α	Attention coefficient
10	Fl	Feature map of l-th layer	30	x^*	Calibrated feature map
11	Fl-1	Feature map of (l-1)-th layer	31	$\phi256$	1x1 convolution with batch normalization
12	$f1 \times 1$	1x1 convolution operation in the gating module	32	ySA1	Output of SAD1
13	Fg	Output of the gating module	33	xl	Low-level feature map
14	$f2 \times 2$	2x2 convolution operation	34	xh	High-level feature map
15	Fcon	Concatenated feature map	35	ϕC	1x1 convolution with batch normalization and C channels
16	cCAM	Channel attention map	36	ySAs	Output of SAD2-4
17	cSAM	Spatial attention map	37	σ	Sigmoid activation function
18	F'	Feature map after cCAM	38	Fav	Average pooling
19	F''	Feature map after cSAM	39	Fmp	Max-pooling
20	F'''	Final output feature map			

Data availability

The dataset used in this study is proprietary and cannot be shared publicly.

References

[1] V.K. Shrivastava, et al., A novel and robust Bayesian approach for segmentation of psoriasis lesions and its risk stratification, *Comput. Methods Progr. Biomed.* 150 (2017) 9–22.

[2] M. Kass, A. Witkin, D. Terzopoulos, Snakes: active contour models, *Int. J. Comput. Vis.* 1 (4) (1988) 321–331.

[3] J.A. Sethian, *Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science*, vol. 3, Cambridge university press, 1999.

[4] J.S. Suri, S. Singh, L. Reden, Fusion of region and boundary/surface-based computer vision and pattern recognition techniques for 2-D and 3-D MR cerebral cortical segmentation (Part-II): a state-of-the-art review, *Pattern Anal. Appl.* 5 (1) (2002) 77–98.

[5] L. Saba, et al., The present and future of deep learning in radiology, *European journal of radiology* 114 (2019) 14–24.

[6] B. Mainak, et al., State-of-the-art review on deep learning in medical imaging, *Frontiers in Bioscience-Landmark* 24 (3) (2019) 380–406.

[7] J.S. Suri, S. Laxminarayan, *PDE and Level Sets*, Springer Science & Business Media, 2002.

[8] J.S. Suri, et al., Shape recovery algorithms using level sets in 2-D/3-D medical imagery: a state-of-the-art review, *IEEE Trans. Inf. Technol. Biomed.* 6 (1) (2002) 8–28.

[9] A. El-Baz, J.S. Suri, *Level Set Method in Medical Imaging Segmentation*, CRC Press, 2019.

[10] J.S. Suri, White matter/gray matter boundary segmentation using geometric snakes: a fuzzy deformable model, in: *International Conference on Advances in Pattern Recognition*, Springer, 2001.

[11] M. Park, et al., ES-UNet: efficient 3D medical image segmentation with enhanced skip connections in 3D UNet, *BMC Med. Imag.* 25 (1) (2025) 327.

[12] V. Kuppili, et al., Extreme learning machine framework for risk stratification of fatty liver disease using ultrasound tissue characterization, *J. Med. Syst.* 41 (10) (2017) 1–20.

[13] U.R. Acharya, et al., Understanding symptomatology of atherosclerotic plaque by image-based tissue characterization, *Comput. Methods Progr. Biomed.* 110 (1) (2013) 66–75.

[14] A. El-Baz, G. Gimel'farb, J.S. Suri, *Stochastic Modeling for Medical Image Analysis*, CRC Press, 2015.

[15] J.S. Suri, R.M. Rangayyan, Breast imaging, mammography, and computer-aided diagnosis of breast cancer, in: *SPIE, SPIE*, Bellingham, WA, USA, 2006.

[16] V.K. Shrivastava, et al., Reliable and accurate psoriasis disease classification in dermatology images using comprehensive feature space in machine learning paradigm, *Expert Syst. Appl.* 42 (15–16) (2015) 6184–6195.

[17] A.S. El-Baz, et al., *Multi Modality state-of-the-art Medical Image Segmentation and Registration Methodologies*, vol. 1, Springer Science & Business Media, 2011 vol. 1.

[18] L. Saba, J.S. Suri, *Multi-Detector CT Imaging Handbook, Two Volume Set*, CRC Press, 2022.

[19] J.M. Sanches, A.F. Laine, J.S. Suri, *Ultrasound Imaging*, Springer, 2012.

[20] R.H. Hashemi, W.G. Bradley, C.J. Lisanti, *MRI: the Basics: the Basics*, Lippincott Williams & Wilkins, 2012.

[21] A.D. Jamthikar, et al., Multiclass machine learning vs. conventional calculators for stroke/CVD risk assessment using carotid plaque predictors with coronary

- angiography scores as gold standard: a 500 participants study, *Int. J. Cardiovasc. Imag.* 37 (4) (2021) 1171–1187.
- [22] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III* 18, Springer, 2015.
 - [23] O. Oktay, et al., Attention u-net: learning where to look for the pancreas, *arXiv preprint arXiv:03999* (2018).
 - [24] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
 - [25] S. Woo, et al., Cbam: convolutional block attention module, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
 - [26] G. Hou, et al., Af-net: a medical image segmentation network based on attention mechanism and feature fusion, *Comput. Mater. Continua (CMC)* 69 (2) (2021) 1877–1891.
 - [27] M. Chen, et al., Multiscale dual attention mechanism for fluid segmentation of optical coherence tomography images, *Appl. Opt.* 60 (23) (2021) 6761–6768.
 - [28] L. Xiao, et al., Enhanced medical image segmentation using U-Net with residual connections and dual attention mechanism, *Eng. Appl. Artif. Intell.* 153 (2025) 110794.
 - [29] L. Xiao, B. Zhou, C. Fan, Automatic brain MRI tumors segmentation based on deep fusion of weak edge and context features, *Artif. Intell. Rev.* 58 (5) (2025) 154.
 - [30] A. Vaswani, et al., Attention is all you need, *Adv. Neural Inf. Process. Syst.* 30 (2017).
 - [31] J. Kim, et al., Hybrid Local-Window-Attention-Assisted U-Net model for multimodal medical-image segmentation, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2025.
 - [32] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning internal representations by error propagation. Institute for Cognitive Science, University of California, San Diego La, 1985.
 - [33] S. Hochreiter, J.J.N.c. Schmidhuber, Long short-term memory 9 (8) (1997) 1735–1780.
 - [34] D. Bahdanau, K. Cho, Y.J.a.p.a. Bengio, Neural Machine Translation by Jointly Learning to Align and Translate, 2014.
 - [35] P. Anderson, et al., Bottom-up and top-down attention for image captioning and visual question answering, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
 - [36] S. Jetley, N.A. Lord, N.J.I. Lee, P. Torr, Learn to Pay Attention, 2018.
 - [37] V. Mnih, et al., Recurrent Models of Visual Attention, vol. 27, 2014.
 - [38] A. Vaswani, et al., Attention is all you need, vol. 30, 2017.
 - [39] O. Oktay, et al., Attention u-net: Learning Where to Look for the Pancreas, 2018.
 - [40] L. Floridi, M.J.M. Chiriatti, Machines, *GPT-3: Its Nature, Scope, Limits, and Consequences*, vol. 30, 2020, pp. 681–694.
 - [41] M.V. Koroteev, BERT: A Review of Applications in Natural Language Processing and Understanding, 2021.
 - [42] K. Han, et al., A survey on vision transformer 45 (1) (2022) 87–110.
 - [43] J. Achiam, et al., Gpt-4 Technical Report, 2023.
 - [44] P. Pan, et al., Multi-scale conv-attention U-Net for medical image segmentation, *Sci. Rep.* 15 (1) (2025) 12041.
 - [45] S. Yan, et al., MDEU-Net: medical image segmentation network based on multi-head multi-scale cross-axis, *Sensors* 25 (9) (2025) 2917.
 - [46] M.A. Al-Ebrahim, Spike-Based attention mechanisms for enhanced medical image segmentation, *Eng. Technol. Appl. Sci. Res.* 15 (5) (2025) 28273–28285.
 - [47] S. Xia, et al., A lightweight neural network for cell segmentation based on attention enhancement, *Information* 16 (4) (2025) 295.
 - [48] A. Hatamizadeh, D. Terzopoulos, A. Myronenko, End-to-end boundary aware networks for medical image segmentation, in: *Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings* 10, Springer, 2019.
 - [49] Z. Qiao, C. Du, RAD-UNet: a residual, Attention-Based, dense UNet for CT sparse reconstruction, *J. Digit. Imag.* 35 (6) (2022) 1748–1758.
 - [50] Y. Li, et al., MHAU-Net: skin lesion segmentation based on multi-scale hybrid residual attention network, *Sensors* 22 (22) (2022) 8701.
 - [51] M. Yin, et al., Cloud and snow detection of remote sensing images based on improved Unet3+, *Sci. Rep.* 12 (1) (2022) 14415.
 - [52] V.K. Singh, et al., HaTU-Net: harmonic attention network for automated ovarian ultrasound quantification in assisted pregnancy, *Diagnostics* 12 (12) (2022) 3213.
 - [53] Y. Wang, J. Wang, P. Guo, Eye-UNet: a UNet-based network with attention mechanism for low-quality human eye image segmentation, *Signal, Image Video Processing* (2022) 1–7.
 - [54] K. He, et al., Deep residual learning for image recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
 - [55] Y. Bengio, P. Simard, P. Frasconi, Learning long-term dependencies with gradient descent is difficult, *IEEE Trans. Neural Network.* 5 (2) (1994) 157–166.
 - [56] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings*, 2010.
 - [57] S. Hochreiter, Untersuchungen Zu Dynamischen Neuronalen Netzen, Technische Universität München, München, Germany, 1991 [Ph. D. dissertation].
 - [58] L.-C. Chen, et al., Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs 40 (4) (2017) 834–848.
 - [59] J. Han, C. Moraga, The influence of the sigmoid function parameters on the speed of backpropagation learning, in: *International Workshop on Artificial Neural Networks*, Springer, 1995.
 - [60] D. Jiang, et al., Automatic supraglacial Lake extraction in Greenland using Sentinel-1 SAR images and attention-based U-Net, *Remote Sens.* 14 (19) (2022) 4998.
 - [61] J. Schlemper, et al., Attention gated networks: learning to leverage salient regions in medical images, *Med. Image Anal.* 53 (2019) 197–207.
 - [62] A. Machireddy, et al., Malignancy detection in prostate multi-parametric MR images using U-net with attention, in: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2020.
 - [63] C. Li, et al., Attention unet++: a nested attention-aware u-net for liver ct image segmentation, in: *2020 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2020.
 - [64] Z. Li, et al., Residual-Attention UNet++: a nested residual-attention U-Net for medical image segmentation, *Applied Sciences* 12 (14) (2022) 7149.
 - [65] C. Li, et al., ANU-Net: Attention-based Nested U-Net to exploit full resolution features for medical image segmentation, *Comput. Graph.* 90 (2020) 11–20.
 - [66] C. Wu, Y. Zou, Z. Yang, U-GAN: generative adversarial networks with U-Net for retinal vessel segmentation, in: *2019 14th International Conference on Computer Science & Education (ICCSE)*, IEEE, 2019.
 - [67] F. Tang, et al., CMU-Net: a strong ConvMixer-based medical ultrasound image segmentation network, *arXiv preprint arXiv:13012* (2022).
 - [68] N. Abraham, N.M. Khan, A novel focal tversky loss function with improved attention u-net for lesion segmentation, in: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, IEEE, 2019.
 - [69] Y.-C. Liu, et al., Cascaded atrous dual attention U-Net for tumor segmentation, *Multimedia tools* 80 (2021) 30007–30031.
 - [70] A. Huang, et al., Attention-VGG16-UNet: a novel deep learning approach for automatic segmentation of the median nerve in ultrasound images, *Quant. Imag. Med. Surg.* 12 (6) (2022) 3138.
 - [71] Z. Su, et al., 3D Attention U-Net: a Solution to CADA-Aneurysm Segmentation Challenge, 2021.
 - [72] T.L.B. Khanh, et al., Enhancing U-Net with spatial-channel attention gate for abnormal tissue segmentation in medical imaging, *Applied Sciences* 10 (17) (2020) 5729.
 - [73] Z. Cheng, et al., Attention V-Net: a residual U-Net with attention Gate block for lung organs at risk segmentation, in: *Proceedings of the 4th International Conference on Computer Science and Application Engineering*, 2020.
 - [74] H. Zhang, H. Zhu, X. Ling, Polar coordinate sampling-based segmentation of overlapping cervical cells using attention U-Net and random walk, *Neurocomputing* 383 (2020) 212–223.
 - [75] D. Lin, et al., RefineU-Net: improved U-Net with progressive global feedbacks and residual attention guided local refinement for medical image segmentation, *Pattern Recognit. Lett.* 138 (2020) 267–275.
 - [76] Z. Wang, L. Zhang, Semantic segmentation of brain MRI based on U-Net network and edge loss, in: *2020 19th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)*, IEEE, 2020.
 - [77] J. Lo, et al., Cross attention squeeze excitation network (CASE-Net) for whole body fetal mri segmentation, *Sensors* 21 (13) (2021) 4490.
 - [78] V. Ashkani Chenarlogh, et al., Clinical target segmentation using a novel deep neural network: double attention Res-U-Net 12 (1) (2022) 6717.
 - [79] J. Wang, et al., EAR-U-Net: EfficientNet and Attention-based Residual U-Net for Automatic Liver Segmentation in CT, 2021.
 - [80] I. Aboussaleh, et al., Efficient U-Net Architecture with multiple encoders and attention mechanism decoders for brain tumor segmentation, *Diagnostics* 13 (5) (2023) 872.
 - [81] S. Agarwala, S. Sharma, B.U. Shankar, A-UNet: attention 3D UNet architecture for multiclass segmentation of Brain Tumor, in: *2022 IEEE Region 10 Symposium (TENSYP)*, IEEE, 2022.
 - [82] Y. Liu, et al., Straw segmentation Algorithm based on modified UNet in complex Farmland environment 66 (1) (2021).
 - [83] Z. Cao, et al., Physics-Informed MTA-UNet: prediction of Thermal Stress and Thermal deformation of satellites, *Aerospace* 9 (10) (2022) 603.
 - [84] S. Jetley, et al., Learn to pay attention, *arXiv preprint arXiv:02391* (2018).
 - [85] T. Shen, et al., Disan: directional self-attention network for rnn/cnn-free language understanding, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.
 - [86] F. Wang, et al., Residual attention network for image classification, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
 - [87] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, *arXiv preprint arXiv:02391* (2014).
 - [88] M.-T. Luong, H. Pham, C.D. Manning, Effective approaches to attention-based neural machine translation, *arXiv preprint arXiv:04025* (2015).
 - [89] Z. Wang, Y. Zou, P.X. Liu, Hybrid dilation and attention residual U-Net for medical image segmentation, *Comput. Biol. Med.* 134 (2021) 104449.
 - [90] X. Ding, et al., CAB U-Net: an end-to-end category attention boosting algorithm for segmentation, *Computerized Medical Imaging Graphics* 84 (2020) 101764.
 - [91] A. Pennisi, et al., Skin lesion area segmentation using attention squeeze U-Net for embedded devices, *J. Digit. Imag.* 35 (5) (2022) 1217–1230.
 - [92] W. Fang, X.-h. Han, Spatial and channel attention modulated network for medical image segmentation, in: *Proceedings of the Asian Conference on Computer Vision*, 2020.

- [93] D. Lin, et al., MLSA-UNet: End-to-End multi-level spatial attention guided UNet for industrial defect segmentation, in: 2022 IEEE International Conference on Image Processing (ICIP), IEEE, 2022.
- [94] D. Zhao, et al., Feature Pyramid U-Net with attention for semantic segmentation of forward-looking sonar images, *Sensors* 22 (21) (2022) 8468.
- [95] Pei, X., Multi-Attention Based Double Unet for Optic Cup and Disc Segmentation (Mad-Unet).
- [96] Z. Wang, Z. Zhang, I. Voiculescu, RAR-U-Net: a residual encoder to attention decoder by residual connections framework for spine segmentation under noisy labels, in: 2021 IEEE International Conference on Image Processing (ICIP), IEEE, 2021.
- [97] Z. Zhong, et al., Full-scale attention network for automated organ segmentation on head and neck CT and MR images, *IET Image Process.* 17 (3) (2023) 660–673.
- [98] Y. Liu, DSNet: a simple yet efficient network with dual-stream attention for lesion segmentation, *arXiv preprint arXiv:16950* (2022).
- [99] Z. Zeng, et al., RIC-Unet: an improved neural network based on Unet for nuclei segmentation in histology images, *IEEE Access* 7 (2019) 21420–21428.
- [100] Z. Huang, et al., GCAUNet: a group cross-channel attention residual UNet for slice based brain tumor segmentation, *Biomedical Signal Processing* 70 (2021) 102958.
- [101] E. Adil, et al., Cascaded U-Net with training wheel attention module for change detection in satellite images, *Remote Sens.* 14 (24) (2022) 6361.
- [102] J. Zhang, et al., Detecting pest-infested Forest damage through multispectral satellite imagery and improved UNet++, *Sensors* 22 (19) (2022) 7440.
- [103] Z. Fan, et al., JAUNet: a U-Shape network with jump attention for semantic segmentation of road scenes, *Applied Sciences* 13 (3) (2023) 1493.
- [104] X. Gao, B. Tong, MRA-UNet: balancing speed and accuracy in road crack segmentation network, *Signal, Image Video Processing* (2022) 1–8.
- [105] J. Lu, et al., A novel hybridoma cell segmentation method based on multi-scale feature fusion and dual attention network, *Electronics* 12 (4) (2023) 979.
- [106] M. Noori, A. Bahri, K. Mohammadi, Attention-guided version of 2D UNet for automatic brain tumor segmentation, in: 2019 9th International Conference on Computer and Knowledge Engineering (ICCKE), IEEE, 2019.
- [107] Y. Xu, et al., A medical image segmentation method based on improved UNet 3+ network, *Diagnostics* 13 (3) (2023) 576.
- [108] L. Xu, et al., Two-Stage UA-GAN for precipitation nowcasting, *Remote Sens.* 14 (23) (2022) 5948.
- [109] Z. Guangzhe, et al., Bilateral u-net semantic segmentation with spatial attention mechanism, *CAAI Transactions on Intelligence Technology* 8 (2) (2022) 297–307.
- [110] C. Guo, et al., Sa-unet: spatial attention u-net for retinal vessel segmentation, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021.
- [111] L. Hong, et al., Qau-Net: quartet attention U-Net for liver and liver-tumor segmentation, in: 2021 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2021.
- [112] I. Delibaşoglu, INC-SA-UNET: spatial ATTENTION INCEPTION UNET FOR AERIAL IMAGES SEGMENTATION, *Comput. Inf.* 40 (6) (2021).
- [113] J. Zhao, et al., Extraction of winter-wheat planting areas using a combination of U-Net and CBAM, *Agronomy* 12 (12) (2022) 2965.
- [114] Y. Zhang, et al., Prostate segmentation using Z-Net, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE, 2019.
- [115] R. Zhou, et al., Deep learning-based measurement of total plaque area in B-mode ultrasound images, *IEEE J. Biomed. Health Inform.* 25 (8) (2021) 2967–2977.
- [116] H. Huang, et al., Unet 3+: a full-scale connected unet for medical image segmentation, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020.
- [117] L. Liu, et al., A survey on U-shaped networks in medical image segmentations, *Neurocomputing* 409 (2020) 244–258.
- [118] N. Siddique, et al., U-net and its variants for medical image segmentation: a review of theory and applications, *IEEE Access* 9 (2021) 82031–82057.
- [119] G. Du, et al., Medical image segmentation based on u-net: a review, *J. Imag. Sci. Technol.* 64 (2) (2020), 20508-1-20508-12.
- [120] G. Samarasinghe, et al., Deep learning for segmentation in radiation therapy planning: a review, *Journal of Medical Imaging Radiation Oncology* 65 (5) (2021) 578–595.
- [121] A. Lagree, et al., A review and comparison of breast tumor cell nuclei segmentation performances using deep convolutional neural networks, *Sci. Rep.* 11 (1) (2021) 1–11.
- [122] S.B. Saadi, et al., Osteolysis: a Literature Review of Basic Science and Potential Computer-based Image Processing Detection Methods, *Computational Intelligence Neuroscience*, 2021, p. 2021.
- [123] N. Singh Punn, S. Agarwal, Modality Specific U-Net Variants for Biomedical Image Segmentation: a Survey, *arXiv e-prints*, 2021 04537 *arXiv*: 2107.
- [124] X.-X. Yin, et al., U-Net-Based medical image segmentation, *J. Healthc. Eng.* 2022 (2022).
- [125] Y. He, et al., Image-to-image translation of label-free molecular vibrational images for a histopathological review using the UNet+/seg-cGAN model, *Biomed. Opt. Express* 13 (4) (2022) 1924–1938.
- [126] J. Wu, et al., A State-of-the-art Survey of U-Net in microscopic image analysis: from simple usage to structure mortification, *arXiv preprint arXiv:06465* (2022).
- [127] R. Zhang, A.C. Chung, EfficientQ: an efficient and accurate post-training neural network quantization method for medical image segmentation, *Med. Image Anal.* 97 (2024) 103277.
- [128] Q. Dou, et al., Unpaired multi-modal segmentation via knowledge distillation, *IEEE Trans. Med. Imag.* 39 (7) (2020) 2415–2425.
- [129] P.K. Rao, S. Chatterjee, S. Sharma, Weight pruning-UNet: weight pruning UNet with depth-wise separable convolutions for semantic segmentation of kidney tumors, *Journal of Medical Signals & Sensors* 12 (2) (2022) 108–113.
- [130] L. Chen, L. Zhao, C.Y.C. Chen, Enhancing adversarial defense for medical image analysis systems with pruning and attention mechanism, *Medical physics* 48 (10) (2021) 6198–6212.
- [131] G. Ma, et al., Attention-based UNet enabled lightweight image semantic communication System over internet of things, *arXiv preprint arXiv:2401.07329* (2024).
- [132] S. Jiang, J. Li, TransCUNet: UNet cross fused transformer for medical image segmentation, *Comput. Biol. Med.* 150 (2022) 106207.
- [133] H. Cao, et al., Swin-unet: Unet-like pure transformer for medical image segmentation, in: European Conference on Computer Vision, Springer, 2022.
- [134] R. Gu, et al., CA-Net: comprehensive attention convolutional neural networks for explainable medical image segmentation, *IEEE Trans. Med. Imag.* 40 (2) (2020) 699–711.
- [135] X. Wang, et al., Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.
- [136] H. Song, et al., RAFF-Net: an improved tongue segmentation algorithm based on residual attention network and multiscale feature fusion, *Digit. Health* 8 (2022), 20552076221136362.
- [137] C. Guo, et al., Channel attention residual u-net for retinal vessel segmentation, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021.
- [138] M. Islam, et al., Brain tumor segmentation and survival prediction using 3D attention UNet, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part I 5, Springer, 2020.
- [139] A. Amer, T. Lambrou, X. Ye, MDA-unet: a multi-scale dilated attention U-net for medical image segmentation, *Applied Sciences* 12 (7) (2022) 3676.
- [140] R. Gandhi, Y. Hong, Mda-net: Multi-dimensional attention-based neural network for 3d image segmentation, in: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), IEEE, 2021.
- [141] Q. Xu, W. Duan, An automatic nuclei image segmentation based on multi-scale split-attention U-Net, in: MICCAI Workshop on Computational Pathology, PMLR, 2021.
- [142] Q. Xu, et al., DCSAU-Net: a deeper and more compact split-attention U-Net for medical image segmentation, *Comput. Biol. Med.* 154 (2023) 106626.
- [143] W. Yuan, et al., DCAU-Net: dense convolutional attention U-Net for segmentation of intracranial aneurysm images, *Visual computing for industry, biomedicine, art* 5 (1) (2022) 1–18.
- [144] P. Jing, et al., Road crack detection using deep neural network based on attention mechanism and residual structure, *IEEE Access* 10 (2022) 120798–120809.
- [145] E. Bourigault, et al., Multimodal PET/CT Tumour Segmentation and Prediction of Progression-free Survival Using a full-scale Unet with Attention, in Head and Neck Tumor Segmentation and Outcome Prediction: Second Challenge, HECKTOR 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings, Springer, 2022, pp. 189–201.
- [146] R. Chang, et al., Automatic extraction of urban impervious surface based on SAH-Unet, *Remote Sens.* 15 (4) (2023) 1042.
- [147] J. Lee, et al., Bilateral attention U-Net with dissimilarity attention Gate for change detection on remote sensing imageries, *Applied Sciences* 13 (4) (2023) 2485.
- [148] J. Li, et al., Semantic segmentation of hyperspectral remote sensing images based on PSE-UNet model, *Sensors* 22 (24) (2022) 9678.
- [149] W. Liu, Y. Sun, Q. Ji, MDAN-UNet: multi-scale and dual attention enhanced nested U-Net architecture for segmentation of optical coherence tomography images, *Algorithms* 13 (3) (2020) 60.