ORIGINAL ARTICLE

DoubleU-NetPlus: a novel attention and context-guided dual U-Net with multi-scale residual feature fusion network for semantic segmentation of medical images

Md. Rayhan Ahmed[1] · Adnan Ferdous Ashrafi[2] · Raihan Uddin Ahmed[3] · Swakkhar Shatabda[1] ·
A. K. M. Muzahidul Islam[1] · Salekul Islam[1]

Abstract
Accurate segmentation of the region of interest in medical images can provide an essential pathway for devising effective treatment plans for life-threatening diseases. It is still challenging for U-Net, and its modern state-of-the-art variants to effectively model the higher-level output feature maps of the convolutional units of the network mostly due to the presence of various scales of the region of interest, the intricacy of context environments, ambiguous boundaries, and multiformity of textures in medical images. In this paper, we exploit multi-contextual features and several attention strategies to increase networks' ability to model discriminative feature representation for more accurate medical image segmentation, and we present a novel dual-stacked U-Net-based architecture named DoubleU-NetPlus. The DoubleU-NetPlus incorporates several architectural modifications. In particular, we integrate EfficientNetB7 as the feature encoder module, a newly designed multi-kernel residual convolution module, and an adaptive feature re-calibrating attention-based atrous spatial pyramid pooling module to progressively and precisely accumulate discriminative multi-scale high-level contextual feature maps and emphasize the salient regions. In addition, we introduce a novel triple attention gate module and a hybrid triple attention module to encourage selective modeling of relevant medical image features. Moreover, to mitigate the gradient vanishing issue while incorporating high-resolution features with deeper spatial details, the standard convolution operation is replaced with the attention-guided residual convolution operations, which enables the model to achieve effective and relevant feature maps from a significantly increased network depth. Empirical results confirm that the proposed model accomplishes superior semantic segmentation performance compared to other state-of-the-art approaches on six publicly available benchmark datasets of diverse modalities. The proposed network achieves a Dice score of 85.17%, 99.34%, 94.30%, 96.40%, 95.76%, and 97.10% on DRIVE, LUNA, BUSI, CVCclinicDB, 2018 DSB, and ISBI 2012 datasets.

Keywords Semantic segmentation · Medical image analysis · Computer-aided diagnosis · Convolutional neural networks · Attention mechanisms · Deep learning · Stacked U-Net · Transfer learning

1 Introduction

Medical imaging is a highly critical element in modern medical practice and biotechnology to undertake numerous diagnostic procedures, from wellness and screening to early diagnosis, clinical analysis, treatment selection, image-guided surgery, and subsequent follow-ups for continuous assessments of the patient's health condition [36]. It has become a crucial resource for physicians to understand and assess the disease. Moreover, it is essential to determine the efficacy of the treatment, allowing clinicians to better analyze a patient by creating a pictorial and functional representation of hidden physiological structures of body parts such as bones, organs, tissue, and blood vessels for clinical examination [4, 52] and evaluate various cellular and molecular events. Noninvasive medical imaging techniques, such as X-ray, computerized tomography (CT), ultrasound, colonoscopy, dermoscopy, microscopy, electrocardiogram (ECG), and magnetic resonance imaging (MRI), can reveal crucial anatomical functionality-related information on diseases and anomalies within the body [39].

Extended author information available on the last page of the article

In recent years, deep learning-based methods have achieved incredible success in many challenging tasks in diverse research domains [1, 2, 31, 51]. Semantic medical image segmentation (MIS) is one of the significant areas of research in medical image analysis. In semantic segmentation, every distinct pixel in the image is assigned a distinct category, thus partitioning an image into a set of non-overlapping regions, which can also be regarded as a dense classification problem [29]. MIS refers to the process of distinguishing specific areas within a 2D or 3D medical image, which can facilitate the clinicians to study only the desired parts or region of interest (ROI) of the multi-modal medical images [29]. It is an essential preliminary step for any computer-aided diagnosis (CAD) system and often plays an integral role in both quantitative and qualitative analysis of medical images [4], such as segmentation of polyps [30, 56], lung region [42, 48], brain tumors [85], retinal blood vessels [83], cell nuclei [80], cell contour [70], and breast ultrasound images [40].

During the past decade, the vast majority of architectures created for semantic segmentation in various applications of computer vision (CV) and medical image analysis are based on deep neural networks (DNNs), such as fully convolutional networks (FCNs), coined by Long et al. [38] or encoder-decoder-based convolutional neural networks (CNNs) such as Seg-Net [6]. The establishment of encoder-decoder-based CNNs achieved promising segmentation performance in CV and medical imaging. Nevertheless, U-Net, proposed by Ronneberger et al. [49], made a significant breakthrough in the MIS task by incorporating the idea of skip connections between each symmetric layer of the encoder and decoder. Primarily, the encoder performs multiple convolutions and pooling operations to capture various representations of images, from low to high-level. It decreases the spatial dimensions of each layer and increases the number of channels. More high-level feature maps, such as objects and various shapes, are captured as the architecture goes deeper. On the contrary, the decoder performs multiple up-sample and concatenation operations, followed by convolution operations to predict the segmented mask. It increases the spatial dimensions while decreasing the channels.

Over the few years, several variants of U-Net followed, such as U-Net++ [85], MultiResU-Net [29], LadderNet [87], Attention U-Net [43], R2U-Net [5], DoubleU-Net [30], CE-Net [22], and KiU-Net [65]. Even though these methods have improved the feature representation to a satisfactory level, they are still constrained by a number of significant drawbacks. Similar scale feature maps with various receptive fields that are generated from the convolution kernel have distinct semantic feature representations. The dimension of the receptive field in the convolutional kernel can affect network performance

[2, 58]. Most of the datasets have images where the ROI is of diverse shapes and sizes, for example, polyps in colonoscopy images. When the receptive field is too large, smaller targets can get disregarded, and on the other side, a smaller receptive field can capture redundant information. Hence, processing the image using convolution kernels with different receptive fields is vital for capturing the global contextual representation of features [22]. Because of the substantial loss of spatial information during encoding, it is usually challenging to reconstruct the details of low-level feature maps such as edges, dots, corners, and lines using orthodox de-convolution operations [35]. The resultant feature maps are sparse, resulting in a reduction in segmentation performance. Moreover, U-Net and its variants also suffer from semantic gaps in feature representations because of longer skip connections present between the corresponding encoder and decoder. Combining the two incompatible representations of feature sets of the encoder and decoder blocks introduces inconsistency in the architecture's learning process. In order to reduce the semantic gaps and loss of spatial information during encoding and improve the high- and low-level fusion of semantic information throughout the network, multiple U-Net-based architectures can be deployed to achieve state-of-the-art (SOTA) segmentation results [30].

The attention mechanism concentrates solely on the most informative feature representations for a specific task without additional supervision, thereby penalizing less informative regions of the image and avoiding using similar feature maps in the network; thus, attention-based networks have recently been widely employed in MIS tasks. The channel-based attention mechanism is one of the most investigated attention mechanisms in literature. It exploits the inter-channel relations of features and focuses on desired object selection by actively re-calibrating each channel's weight [23]. Hu et al. [28] initially presented the idea of channel attention and introduced SE-Net architecture. SE-Net utilizes the global average pooling mechanism to capture the global representations of contextual features. However, a simple global average pooling mechanism can fail to extract complex high-level intra-channel feature representations [20]. The spatial attention mechanism mainly focuses on relevant spatial regions of informative features. Nevertheless, integration of only SE-Net's attention has been found to be inadequate and sub-optimal in many MIS tasks [71]. Woo et al. [71] suggested the concept of a convolutional block attention module (CBAM), which is a sequential combination of these two attention mechanisms and can bring effective results for many CNN-based tasks. Oktay et al. [43] introduced a low-cost, lightweight attention gate mechanism to focus mainly on the selected ROIs while suppressing feature activations in non-ROIs. Recently various transfer learning techniques

have been applied to the task of MIS due to their robustness and quick convergence mechanism [22, 30, 68]. It allows the pre-trained weights from one task to be utilized in different but related tasks.

In this paper, we extend and significantly improve the SOTA DoubleU-Net [30] architecture and propose a robust novel architecture that can effectively perform the MIS tasks of multi-modal domains by modeling global contextual information and high-level multi-scale semantic feature representations of pixels of varying receptive fields. EfficientNetB7 [61] architecture is adopted through transfer learning as our backbone encoder module for extracting effective feature information. We incorporate a novel triple attention gate (TAG) mechanism in every skip connection to attend to selective inputs with high relevancy to the target region. To reduce the semantic gap issues of the skip connections of U-Net [49], DoubleU-Net [30], and other similar variants, we incorporate attention-guided residual (AG-Residual) convolution operations instead of regular convolutions. We also design a multi-kernel residual convolution (MKRC) module to acquire high-level global contextual features. The MKRC block extracts fine-grained contextual information of higher levels from images with various levels of receptive fields such as 1×1, 3×3, 5×5, and 7×7. The receptive field of a CNN usually refers to the size of the kernel. The generated feature maps from the MKRC block are then passed through the newly designed squeeze and excitation-based atrous spatial pyramid pooling (SE-ASPP) module [14] to extract high-resolution relevant feature maps for effective learning of the proposed model. In addition, inspired by the CBAM architecture [71], we also integrate a hybrid triple attention module (TAM), which performs features refinement through parallel execution of spatial attention, modified channel-based attention mechanism, and squeeze and excitation-based attention to capturing relevant spatial regions of the higher-level global contextual features and inter-dependencies among different channels, respectively.

Overall, the main contributions of this work can be summarized as follows:

- A robust EfficientNetB7 encoder backbone-based segmentation framework, referred to as DoubleU-NetPlus, is proposed to enhance the semantic segmentation performance for biomedical images.
- A newly proposed multi-kernel residual convolution module, which expands the field of view representation of heterogeneous, semantic global contextual features at different scales.
- A modified hybrid triple attention module, which performs an aggregation of spatial and channel-based attention and squeeze and excitation-based attention, thus, improves the channel inter-dependencies and

inter-spatial relationships of the high-level feature maps.
- A novel lightweight triple attention gate module is integrated at the decoder side of each network to highlight salient features from the skip connections.
- Embedding of features re-calibration through squeeze and excitation operation in the attention-based atrous spatial pyramid pooling mechanism.
- We demonstrate the effectiveness of the proposed DoubleU-NetPlus architecture on six publicly available benchmark datasets of different modalities, and comparative analysis exhibits that the proposed method outperformed several SOTA medical image segmentation methods.

2 Related works

This section provides a brief summary of the research pertaining to MIS techniques, including context-aware segmentation, attention-guided segmentation, and stacked multi-U-Net techniques.

2.1 Context-aware segmentation

Contextual information from multiple levels of a network plays a significant role in the performance of any CNN-based MIS model. Xie et al. [76] proposed a context hierarchical integrated network (CHI-Net), which introduced a dense dilated convolution module for gathering features from four cascaded branches of hybrid dilated convolutions. The authors also introduced a stacked residual pooling module that uses multiple effective fields of view. Residual dilated convolution was utilized in the encoder part of the network to capture multi-level hierarchical features. Gu et al. [22] used a context encoder network (CE-Net) that utilizes a pre-trained ResNet34 as the encoder module. The authors integrate a context extractor module consisting of a dense atrous convolution block and a residual multi-kernel pooling block. Al-masni and Kim [4] applied a contextual multi-scale multi-level network (CMM-Net) by fusing the global contextual features of different spatial scales in the encoding part of the U-Net. The authors also used a dilated convolution module that expanded the receptive field with different rates depending on feature maps network sizes.

Xiao et al. [75] introduced a deep residual contextual and sub-pixel convolution network (RC-SPCNet) for the segmentation of neuronal structure. The encoding section of the U-Net included residual-convolution blocks along with summation-based skip connections, and the decoding section was deployed with sub-pixel convolutional layers.

Lifted multi-cut was used for optimizing the output for reconstruction results. Lou et al. [40] introduced an inverted residual pyramid block and a context-aware fusion block in a new U-Net architecture. The authors deployed a multi-level context refinement network (MCRNet) using these two context refinement blocks into a U-net architecture in a multi-level manner. In another study, Wu et al. [72] proposed a new U-Net architecture comprising three new modules: a scale-aware feature aggregation module, an adaptive feature fusion, and a multi-level semantic supervision module.

Recently various transformer-based architectures have been effectively used in the MIS task too. By modeling global context-based features effectively, architectures like Swin-UNet [10], Ds-TransUNet [34], and UNETR [25] achieved SOTA results on MIS tasks of diverse modalities.

In all of the studies discussed above, the authors tried to extract multi-scale representations to reduce gaps in semantics between the encoder and decoder features. Although, these readjustments in many a case introduced over-fitting problems [81], which resulted in not so much significant rise in evaluation metrics.

2.2 Attention-guided segmentation

Over the years, with the successful application of many computer vision-oriented tasks, various attention mechanisms have been increasingly applied to the field of MIS. Wang et al. [68] proposed an iterative edge attention network (EANet) where the authors integrated the edge-attention preservation (EAP) module along with a dynamic scale-aware context module. The authors employ the VGG-19 [54] pre-trained architecture as the feature encoder. The EAP module captures edge-related attention information such as background noise and shape by preserving the low-level local edge features. The gated convolutional blocks (GCB) interleaved with some residual blocks in the EAP module allow the edge stream to solely analyze boundary-related data.

Zhao et al. [84] proposed an MIS architecture where the authors apply spatial and squeeze and excitation networks (SE-Net) to focus mainly on the initial low-level feature maps and channel inter-dependencies in the high-level feature maps in the bottleneck part of the network. Wang et al. [67] incorporate the SE attention mechanism in the encoder part of the network to adaptively extract the feature maps and the ASPP module to capture the context-based semantic information from the extracted feature maps at multiple scales. SE-Net is also incorporated by Li et al. [33], where the authors use Res2Net [19] as the encoder backbone. The extracted features are grouped by channels, and convolution operations are performed on each group separately. SE-Net is integrated to learn the

relationship between groups and re-calibrate the channel weights to focus on the target object.

Gao et al. [18] proposed a multi-scale fused network that employs two attention mechanisms, additive channel attention and additive spatial attention in the skip connections, which utilize high-level features to prune the responses of low-level features in both channel and spatial dimensions. It improves the learning of the superior spatial relationship between adjacent pixels and inter-dependencies between channels. Yeung et al. [79] proposed an attention-gated U-Net architecture that employs a new attention module named focus gate and combines spatial and channel-based attention with a focal parameter to regulate the degree of background suppression. The focus gate utilizes the gating signal to refine incoming signals from the encoding network as long-range skip connections, indicating selected image features and regions included in the decoding network.

Tomar et al. [63] introduced a new attention-based mechanism named FANet, which combines the feature maps from the current training epoch with the prior epoch mask. The prior epoch mask provides hard attention to the learned feature maps at different convolutional layers. Han et al. [24], in their proposed ConvUNeXt architecture, utilized the ConvNeXt [37] as the encoder backbone along with the attention gate mechanism in every skip connection. Tong et al. [64] also utilized the lightweight attention gate mechanism in the decoder part of the network. The feature map generated by the attention gate module is processed by the channel and spatial attention modules in parallel, whose outputs are combined to produce the final feature maps.

Though all the aforementioned attention-based methods achieved reasonable performance in the MIS tasks, they still face challenges in achieving SOTA segmentation performance in terms of diverse shapes, intricate textures, and subjects, especially in the breast ultrasound image and retinal modalities.

2.3 Stacking/cascading of multiple U-Nets

Another popular method explored by researchers to improve feature representation of segmentation tasks is to stack multiple U-Net architectures together in a k-cascading U-Net format, where k refers to the number of sub-U-Nets [47, 82]. For example, DoubleU-Net [30], with two U-Net architectures stacked on top of each other. Ghosh et al. [21] proposed the idea of incorporating dilated stacked U-Nets for semantic scene segmentation. In another work, Ding et al. [16] utilize a series of U-Nets stacked together for brain tumor segmentation. In addition, a multi-level nested U-net structure with encoders and decoders comprised of U-Net structured modules has been

constructed [47] for salient object detection and segmentation. Furthermore, W-shaped networks have been established in recent years. W-Net [74] functions by concatenating two U-Nets into an autoencoder format, one for encoding and one for decoding, and achieves satisfactory results in unsupervised image segmentation tasks.

All of the above-mentioned architectures connect two or more U-Nets together and can therefore extract a separate group of features using the same set of original features. However, the challenge is that the same features may be extracted repeatedly, which can degrade the network's efficiency [82].

3 Proposed method

In this section, we describe the architecture of the proposed segmentation network and the details of the constituent modules. Firstly, the architecture of the DoubleU-Net [30] model is briefly described, and then we elaborately describe the proposed architecture and the incorporated modules in it. The proposed architecture is demonstrated in Fig. 1.

3.1 Overview of DoubleU-Net architecture

DoubleU-Net [30] is an encoder–decoder architecture comprising two U-Net-like networks stacked on top of each other. There are two encoders and two decoders in the DoubleU-Net architecture. In the first U-Net architecture, VGG-19 [54] is incorporated as the backbone of the first encoder, which is pre-trained on ImageNet [32]. The decoder of the first U-Net architecture is built by performing the up-sampling of the feature maps, then concatenating with the corresponding skip connections, and lastly, two regular convolution operations of 3×3 followed by batch normalization, ReLu, and squeeze and excitation operation. In order to utilize more high-level semantic information efficiently, the authors placed the second U-Net at the bottom of the first U-Net. The encoder of the second U-Net is formed by performing consecutive convolution and max-pooling operations. The decoder of the second U-Net is similar to the decoder of the first U-Net. The results generated by the DoubleU-Net architecture outperformed several MIS algorithms by a significant margin in four benchmark datasets. Despite achieving significant performances, DoubleU-Net lacks effectiveness in the skip connections of the network [50], limiting the precise flow of information throughout the network. Moreover, it does not fully exploit the high-level feature maps from varying receptive fields, which can increase the results further. A further shortcoming of DoubleU-Net is its outdated VGG-19 encoder backbone, which can be replaced by a more recently proposed deeper architecture like EfficientNetB7 [24]. Hence, we select DoubleU-Net as our basic architecture for further enhancement.

3.2 Overview of the proposed DoubleU-NetPlus architecture

We performed enhancements in both the networks of the DoubleU-Net architecture by deploying the EfficientNetB7 architecture as the encoder one backbone for extracting multi-scale information. In all the skip connections, we employ a novel triple attention gate (TAG) module to selectively attend to the significantly relevant features in the decoder while suppressing irrelevant feature representations. Compared to high-level feature information, low-level feature information tends to contribute less to network performance and use a lot of computational resources, as pointed out by [55, 73]. As demonstrated in Fig. 1, to capture more effective multi-scale high-level contextual encoder information and pass it to be decoded by the decoder in the bottleneck/bridge of each encoder-decoder network, we design and embed the multi-kernel residual convolution (MKRC) module, modified squeeze and excitation-based atrous spatial pyramid pooling (SE-ASPP) module, and triple attention module (TAM) sequentially. Deeper networks considerably enhance the performance of the model. However, an increase in the depth of the network might occur in vanishing or exploding gradient problems [26, 67]. In order to address this issue and reduce the semantic gaps between the feature representations of the encoder and decoder, we utilize shortcut connections between layers in the residual learning paradigm. We have performed attention-guided residual (AG-Residual) convolution operations (see Fig. 2) in the encoder of the second network and decoders of both networks. The motivation behind deploying two multi-contextual attention-guided residual U-Net architectures is that the output feature maps of network one are not fully explored [82]. We can enhance it by capturing the unexplored high-level multi-contextual information from the generated output feature maps of network one by multiplying it with the original input image and processing them together again in the second network to capture more semantic information.

3.3 Encoder and decoder

The encoder portion of a U-Net is responsible for condensing the spatial information by each level. While it does so, the number of inputs halves, and the number of channels doubles. Consequently, we are left with highly condensed feature information that needs to be passed on to be decoded by the following levels. In our proposed DoubleU-NetPlus architecture, we utilize the EfficientNetB7 pre-
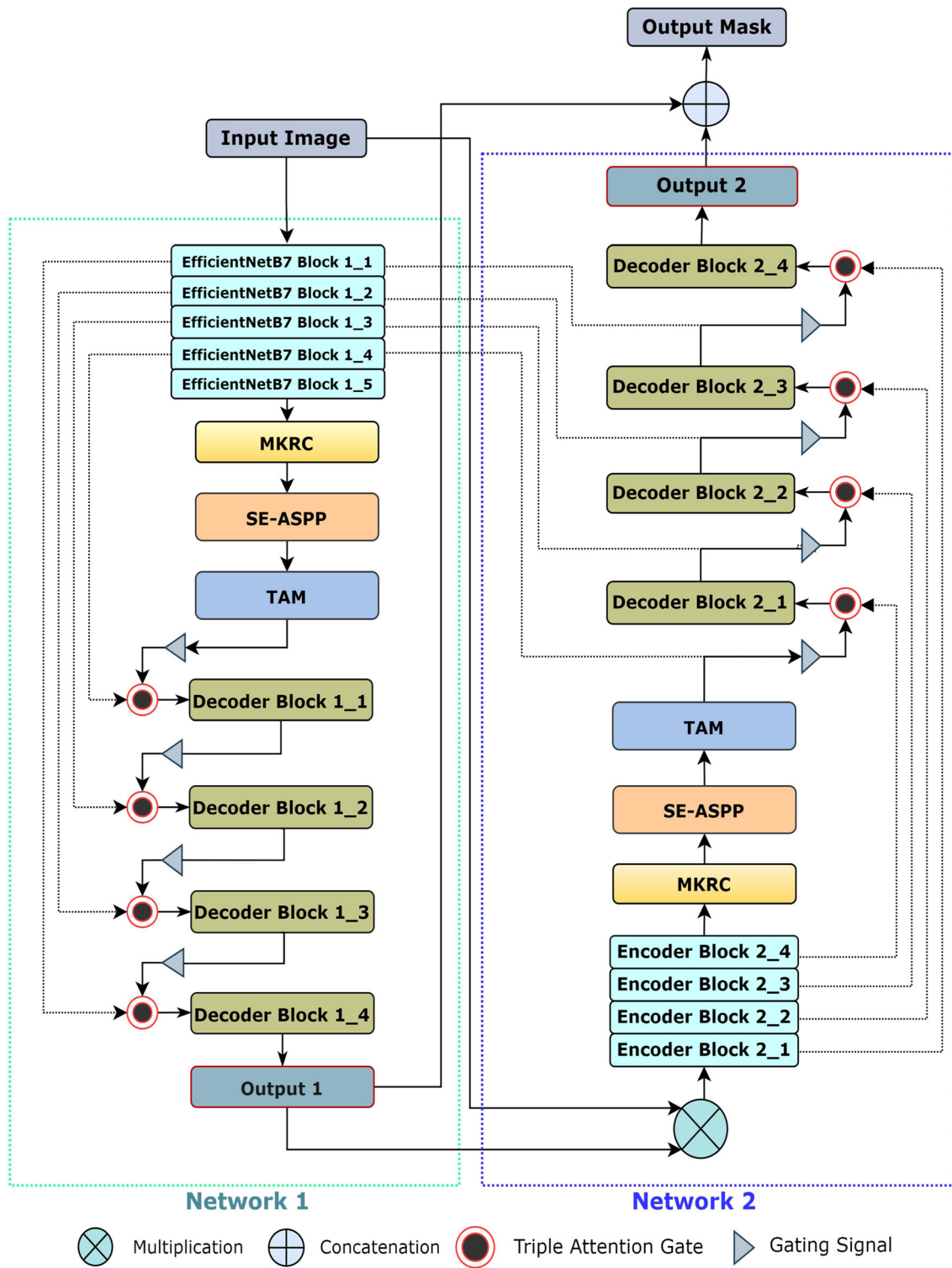
Fig. 1 Composition of the proposed DoubleU-NetPlus architecture

trained architecture as the backbone for the encoder of network one using the transfer learning method, whereas the encoder in network two is built by performing two residual convolutions of 3×3 followed by spatial and

channel attentions. In the first encoder, we chose the EfficientNetB7 architecture mainly because of its higher accuracy and increased network depth. The deployment of EfficientNetB7 as the encoder of the first network gives the
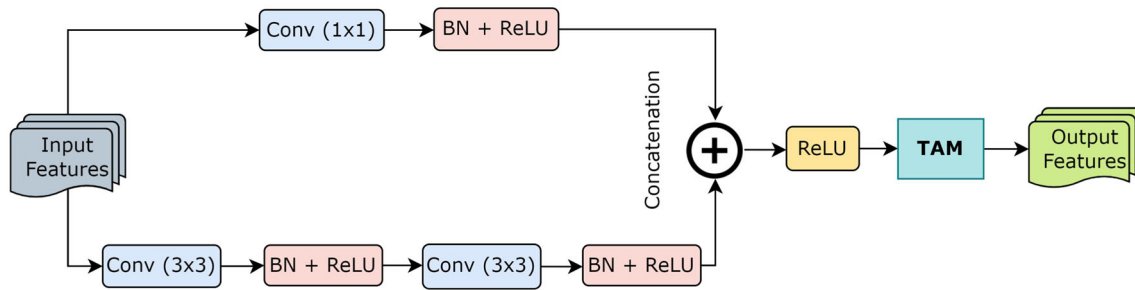
Fig. 2 Composition of the attention gated (AG)-residual convolution module

network effective feature extraction capability that the decoder of the first network can employ to generate extremely precise segmentation maps [53]. EfficientNetB7 implements a mobile inverted bottleneck convolution with an injected SE-Net [28] block, which can attend to relevant features. By utilizing shortcuts directly between bottlenecks, which connect a significantly less number of channels than expansion layers, and depth-wise separable convolution, which effectively reduces computing cost compared to traditional layers. It performs more effectively by uniformly scaling the network's resolution, depth, and width, resulting in improved performance. Hence, deploying an EfficientNetB7 encoder enables us to have a contracting path that is significantly deeper and can perform effective contextual feature extraction of medical images. Each encoder block of the second encoder executes AG-residual convolution operations, as illustrated in Fig. 2. The AG-residual convolution module performs two 3×3 convolution operations, each of which is followed by batch normalization and ReLU. The batch normalization decreases the internal covariant shift and regularizes the model [30], while ReLU introduces nonlinearity to the architecture. A shortcut residual connection is added with a 1×1 convolution of the input features to provide an identity mapping of features, followed by batch normalization and ReLU operations. Features from the 3×3 convolution operations and 1×1 shortcut connection are concatenated, followed by another ReLU operation. The generated feature maps are then passed to the TAM module, which performs both spatial and channel-based attention as well as squeeze and excitation-based attention on the features to focus more on the relevant feature maps. Then we perform a max-pooling operation with a 2×2 window and stride of 2×2 to reduce the spatial dimension of the feature maps.

As shown in Fig. 1, the architecture has two decoders, one in each network. Each input feature is passed to the gating signal module, which captures high-level feature representations from the immediate lower part of the network. Then, each block in the decoder applies a 2×2 up-sampling of bi-liner interpolation to each input feature,

hence doubling the dimension of the input feature maps. The generated feature maps are then passed to the attention gate module, which takes the skip connections, and the gating signal as inputs and performs additive soft attention on these two feature maps, and the network learns to attend to the desired ROI while suppressing feature activation in irrelevant areas as the training proceeds. Then, we concatenate the up-sampled feature maps with the output feature maps of the attention gates. The concatenated feature maps are then passed to the AG-residual module for attention-based convolution operations. Every skip connection in the proposed model passes through the attention gate. In the first decoder, we only employ attention-gated-skip connections from the first encoder of network one. However, in the second decoder, we use attention-gated-skip connections from both the encoders from networks one and two. This procedure maintains spatial resolution and improves the output feature maps' quality without focusing on irrelevant regions. Similar to the DoubleU-Net architecture [30], the final step is applying a convolution layer with a sigmoid activation function to construct the mask for the modified U-Net.

3.4 Multi-kernel residual convolution module

One of the challenges in MIS is the larger variation in the size and shape of an object in the medical image. Hence, to achieve effective results in the MIS task, extracting high-level multi-scale contextual features through different receptive fields is necessary. In our proposed architecture, we applied an inception architecture [60] inspired multi-kernel residual convolution (MKRC) module in both of the bottlenecks of networks one and two, which helps reduce saturation and degradation in the learning gradient. The proposed MKRC module is demonstrated in Fig. 3. The MKRC module expands the field of view representation of heterogeneous features for more effective and robust learning of the model. The module consists of multiple parallel convolution layers with different kernel sizes of (1×1), (3×3), (5×5), and (7×7), respectively. Increasing the kernel size in the convolution layers enables
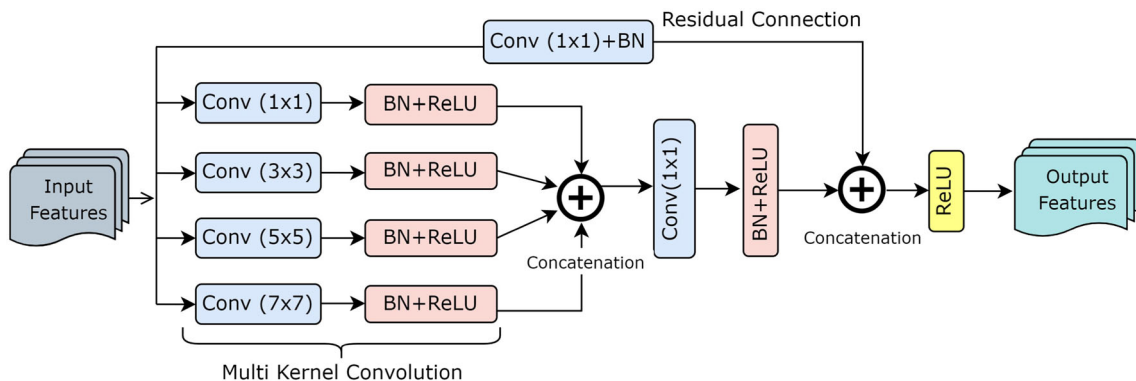
Fig. 3 Composition of the multi-kernel residual convolution (MKRC) module

the networks to extract a more robust feature representation from multi-scale receptive fields, causing them to modulate the learning of features differently for each block. The next step after each convolution layer is a batch normalization layer and a ReLU activation function. After that, all four feature maps are concatenated together, which leaves us with information on every relevant receptive field. Next, we feed the concatenated feature maps to a (1×1) convolution followed by batch normalization and ReLU. Next, we integrate a residual shortcut connection, also known as identity mapping [27], passed through a (1×1) convolution and batch normalization and perform concatenation with the previously generated feature maps. An effective identity mapping through a (1×1) convolution in lesser residual settings can ensure smooth propagation of information in a network with reduced overfitting. A ReLU activation is performed next. The resulting feature maps are then processed through a modified SE-ASPP module that expands the field-of-view representation of features to encompass a broader context.

3.5 Squeeze and excitation-based atrous spatial pyramid pooling module

Atrous spatial pyramid pooling (ASPP) introduced by Chen et al. [14] allows us to effectively enlarge the filters' field of view to include multi-scale contextual representation of semantic features by parallel atrous convolution layers with different dilation rates. It can efficiently mitigate the issue of reduced spatial resolution resulting from repeated downsampling in the encoder [67]. We modify the ASPP module and propose a new SE-ASPP module by embedding the squeeze and excitation networks (SE-Net) to the increased and enlarged field of view of the convolution filters. The structure of the SE-ASPP module is demonstrated in Fig. 4. We have utilized a deeper set of dilated convolutions in the SE-ASPP module in order to capture more robust and expanded representations of features from the MKRC module. The dilation rates utilized in the seven parallel
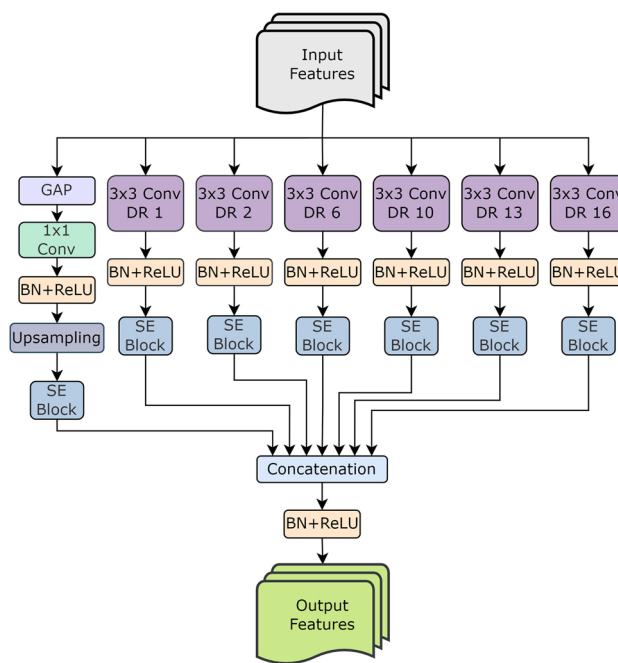

Fig. 4 Composition of the squeeze and excitation-based atrous spatial pyramid pooling (SE-ASPP) module

convolution layers of the SE-ASPP module are 1, 1, 2, 6, 10, 13, and 16, respectively. We apply the squeeze and excitation network to effectively re-calibrate and refine the acquired features through different dilation rates. All the feature maps from the SE-Net modules of each branch of the SE-ASPP network are concatenated together, and a (1×1) convolution operation is performed on the concatenated feature maps, followed by batch normalization and a ReLU activation function. The SE-ASPP module captures efficient and relevant semantic information at multi scales. The generated feature maps are then passed to the hybrid triple attention module (TAM) for further processing.

3.6 Hybrid triple attention module

The hybrid triple attention module (TAM) performs effective attention-based feature refinement and extends the concepts introduced by CBAM [71] and Focus U-Net [79]. As shown in Fig. 5, it performs a feature fusion through parallel processing of squeeze and excitation networks [28], modified channel-based attention, and spatial attention mechanisms. We utilize these attention mechanisms to fully explore the high-level inter-spatial relationship of relevant features and effective inter-channel relationships. By adjusting the weight of each channel, SE-Net offers channel-based attention that can improve the channel inter-dependencies and can be seen as an object selection process while suppressing noise. However, SE-Net performs only global average pooling operations to perform channel-based attention. Later, CBAM [71]

suggested that these features could be sub-optimal and suggested using max pooling operations for modeling improved channel inter-dependencies. As illustrated in Fig. 5, to achieve effective channel-based attention, we extend the ideas of CBAM and employ initial global average pooling and global max pooling operations along the channel axis, followed by concatenation and sigmoid activation to generate efficient feature descriptor that helps to determine where to highlight or suppress along the channel axis. Through the spatial attention mechanism, the architecture focuses on the location of high-level feature maps of the target regions. In conjunction with channel-based attention, spatial attention module aggregate features along the channel axis [28, 71, 79]. We utilized the CBAM implementation of spatial attention by establishing two distinct channel contexts using average and max pooling along the channel axis, following spatial re-calibration

Fig. 5 Composition of the triple attention module (TAM)
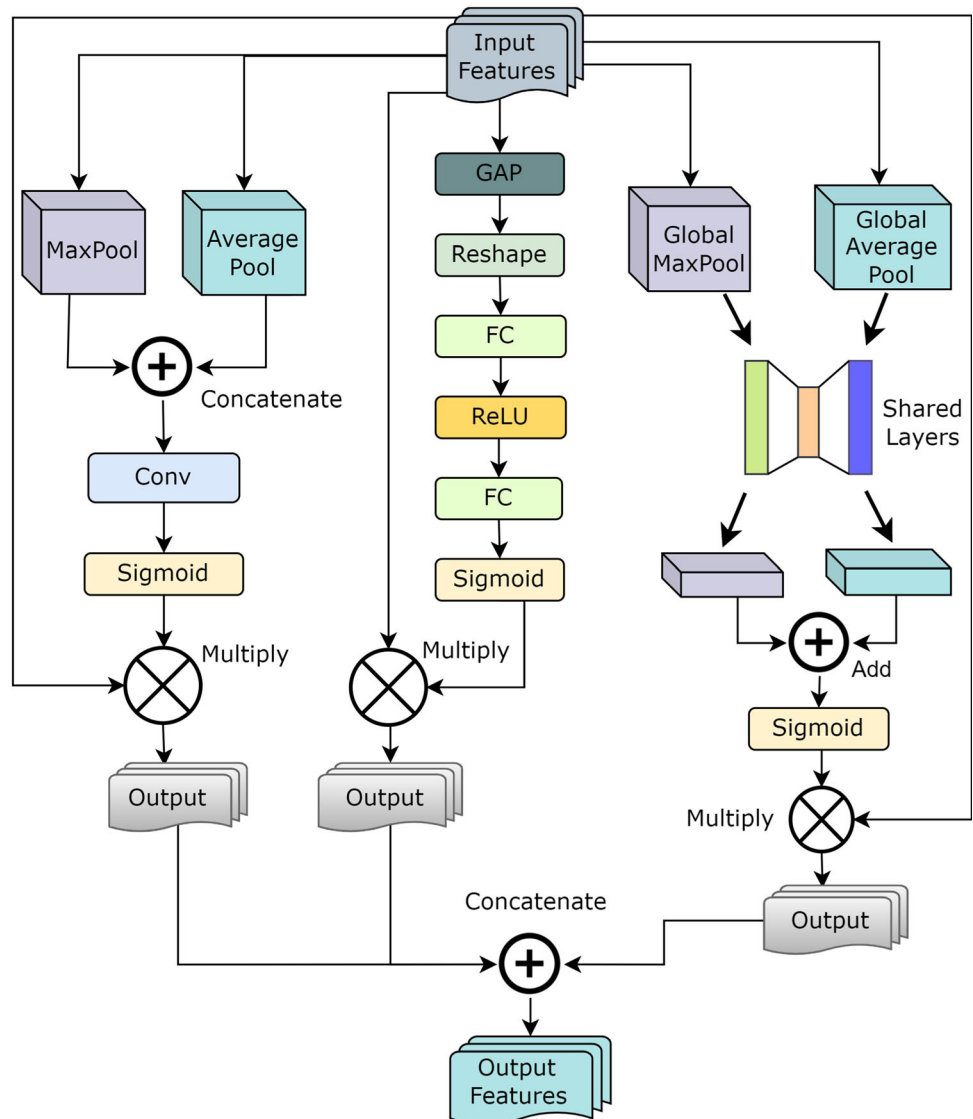
using a kernel of size 7. Similar to modified channel-based attention, we experimented by incorporating initial global average pooling and global max pooling operations in the spatial attention module; however, the performance did not improve, and hence opted to use the original implementation of CBAM.

3.7 Triple attention gate module

Having introduced the SE-Net, channel-based attention, and spatial attention modules in the previous subsection, we describe the structure of the triple attention gate (TAG) module. Due to the lightweight design of the attention gate module, it significantly improves the model's representation ability without significantly increasing the computing cost or the number of model parameters [43]. Here, similar to the attention gate and focus gate [79] modules, we introduce a novel triple attention-gated deep neural network named the TAG module, which performs parallel implementation of channel attention, spatial attention, and squeeze and excitation-based attention mechanisms into a single TAG module to encourage selective learning of efficient, relevant features. The TAG module takes two inputs, as shown in Fig. 6; one is the gating signal from the one-step lower levels, which has a better representation of features such as edges, texture, and dots through training, and the other is the corresponding skip connection at that level, having a better representation of the spatial information. First, the gating signal and skip connection are resized to matching dimensions, and then they are combined through element-wise addition followed by nonlinear activation (ReLu) and create attention coefficients. After that, the attention coefficients are passed through the channel, spatial, squeeze, and excitation-based attention modules and are then concatenated together to produce

effective refinement of the relevant features. Next, we perform a 2×2 up-sampling operation to match the dimensions from the output of the 1×1 convolutions, followed by sigmoid operations, and 2×2 up-sampling performed on the output of the previously mentioned nonlinear activation function. The aligned weights get larger, and the unaligned weights get relatively smaller. The spatial contextual information of ROIs is captured by concatenating the original skip connection by the generated attention coefficients. Hence, the vector gets scaled based on its relevance.

4 Experimental analysis

4.1 Datasets

This section briefly describes all the utilized datasets in this study. For the evaluation of the proposed model, we have utilized six datasets of different modalities, namely BUSI, CVCclinicDB, Drive, ISBI 2012, 2018 DSB, and LUNA. A representative image and the corresponding mask from each of the datasets are shown in Fig. 7.

4.1.1 DRIVE

A diabetic retinopathy screening program in the Netherlands provided the data used to create the Digital Retinal Images for Vessel Extraction (DRIVE) dataset facilitating retinal vessel segmentation as described in [57]. A Canon CR5 non-mydriatic 3CCD camera with a 45-degree field of view (FOV) was used for image acquisition, and there is a total of 40, 8-bit per color channel images with a resolution of 768×584 pixels.
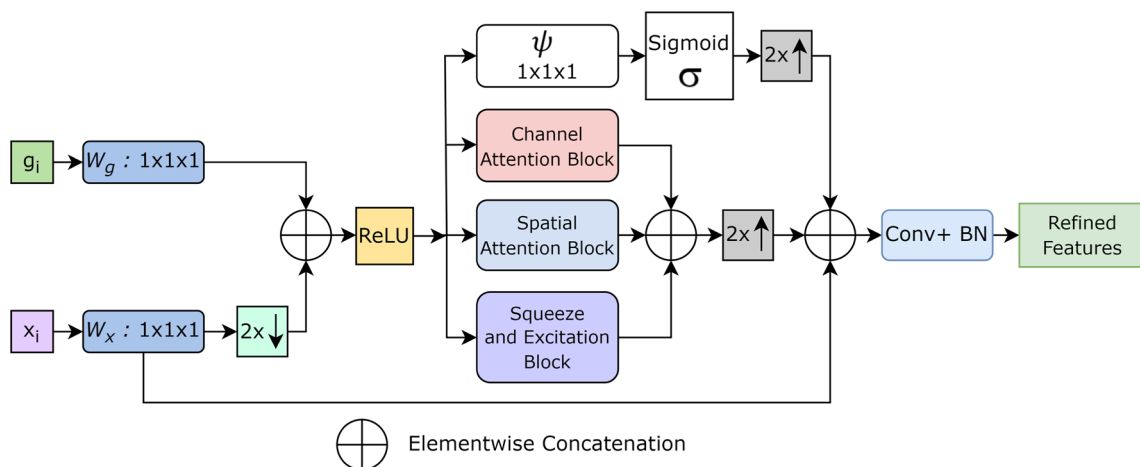

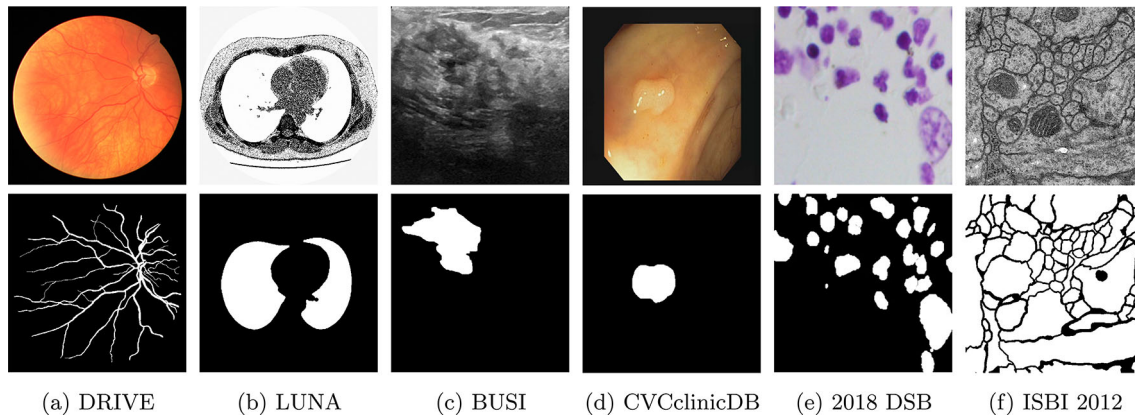**Fig. 6** Composition of the triple attention gate (TAG) module

Fig. 7 Input images and their corresponding segmentation masks in the dataset. Sample images and their masks of DRIVE, LUNA, BUSI, CVCclinicDB, 2018 DSB and ISBI 2012 can be found in (**a**, **b**, **c**, **d**, **e**), and (**f**), respectively

4.1.2 Lung segmentation

Based on the computed tomography (CT) image modality, lung segmentation from CT images is available in the lung nodule analysis (LUNA) competition [41]. This dataset contains 267 2D CT images with full annotations of the labeled lung images provided by experts in the medical sector. The size of each image is 512×512 pixels.

4.1.3 Breast ultrasound image

Utilizing LOGIQ E9 ultrasound system-guided scanning, the breast ultrasound image (BUSI) dataset was created from images collected from 600 females aged between 25 and 75 years old [3]. The dataset contains seven hundred eighty images with an average image size of 500×500 pixels in three distinct categories: benign, normal, and malignant. The ground truth for each image was generated using MATLAB.

4.1.4 CVCclinicDB

The CVCclinicDB dataset contains image frames extracted from colonoscopy videos, using Window Median Depth of Valley (WM-Dova) methodology as mentioned in [7]. From a collection of twenty-nine video sequences, 612 still image frames were extracted for polyp detection. Each image is of the size 384×288, while the corresponding ground truth image is presented as the segmentation mask of the polyps.

4.1.5 2018 data science bowl

The 2018 data science bowl (DSB) dataset was created as a challenge for generic segmentation of nuclei of cells in a diverse set of stained two-dimensional (2D) microscopic images [9]. The training set contains 670 images from both bright-stained and fluorescence modalities of microscopic images with sizes $256 \times 256 \times 3$. In addition to the images captured under various lighting conditions, corresponding annotations (segmentation masks) for each image are also provided to be used as ground truth.

4.1.6 ISBI 2012

The ISBI 2012 dataset, introduced in [11], is comprised of transmission electron microscopy (TEM) images of *Drosophila* larval brain for the purpose of analyzing the structural aspect of neural micro-circuitry. The training data are comprised of 30 TEM 512×512 serial section images of the first instar larval *Dorsophila* brain using TrakEm2 [12] software. The corresponding labels of each image were produced by an expert neuro-anatomist for the purpose of segmentation (Table 1).

4.1.7 Preprocessing and data augmentation

In our experiments, we used several augmentation techniques to ensure that over-fitting does not occur for a small number of samples present in the datasets. To ensure efficient, robust learning of the proposed model in five datasets, namely CVCclincDB, 2018 DSB, BUSI, and LUNA, we employed a total of thirteen data augmentation techniques, including two variations of random rotations, grid distortion, horizontal and vertical flips, transpose, a composition of vertical flip and random rotation, random brightness, random contrast, random brightness contrast, random gamma, hue-saturation contrast, and RGB shifting to increase the image variability during the training process. For the DRIVE and ISBI 2012 datasets, we employed a total of twenty-two data augmentation techniques, that includes the techniques mentioned above, as well as CLAHE, FancyPCA, and Gaussian noise injection. It

Table 1 Overview of the datasets employed in our experiments

Dataset	#Training samples	Modality	Original shape	#Samples after augmentation
DRIVE	20	Retinal	565×684	368
LUNA	267	Computed Tomography	512×512	2992
BUSI	647	Ultrasound	558×462	7080
CVCclinicDB	612	Colonoscopy	384×288	6856
2018 DSB	670	Brightfield & Fluorescence	256×256	7504
ISBI 2012	30	Electron Microscopy	512×512	528

should be noted that the original DoubleU-Net architecture employed a total of twenty-five augmentation types of a single image mask pair. After the data augmentation process, the augmented RGB images were compressed to 256×256 to prepare them for fitting into the models. It should also be noted that the original images were also resized and incorporated into the training dataset.

4.2 Training setup and experimental metrics

In order to train the models, the augmented dataset was divided using an 80:10:10 ratio, i.e., 80% of the images were used for composing the training dataset, 10% for the testing, and the rest 10% for the validation dataset. We initialize the pre-trained weights of EfficientNetB7 architecture, and the batch size was set at 4. The learning rate starts from 0.0001, and the learning rate is reduced by a factor of 0.1, with patience of 10. We fed 2D images of size 256×256 as input for the proposed network. Our system was implemented using Tesla P100-PCIE GPU with 16 GB RAM and a Tensorflow backend. The total number of trainable parameters of the proposed model is 22.4 million. We incorporated a hybrid loss function by adding the binary cross-entropy loss (Loss_{BCE}) and Dice loss ($\text{Loss}_{\text{Dice}}$) [59], offering smooth gradient flow and handling of the class imbalance problems [8]. The hybrid loss function can be defined as:

$$\text{Loss}_{\text{Hybrid}} = \text{Loss}_{\text{BCE}} + \text{Loss}_{\text{Dice}} \tag{1}$$

$$\begin{aligned} \text{Loss}_{\text{BCE}} &= - \sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \\ &= -(y \log(p) + (1-y) \log(1-p)) \end{aligned} \tag{2}$$

The Loss_{BCE} specified in Eq. (2) can be defined in terms of the number of classes M, the natural log, binary indicator (0 or 1) y, class label c the correct classification for observation o, and p is the predicted probability observation o is of class c.

$$\text{Loss}_{\text{Dice}} = 1 - \frac{2 \sum_{i=1}^{N} p_i g_i + \epsilon}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2 + \epsilon} \tag{3}$$

The Dice coefficient between the prediction samples p and the mask g can be defined as given in Eq. (3). Here, ϵ is a constant added to avoid the divide by zero error.

4.2.1 Precision and recall

True positive (TP) outcomes are the number of samples that were correctly classified as the mask, and false positives (FP) are the number of samples that were falsely predicted as part of the mask region. On the other hand, the true negatives (TN) are the number of samples that are correctly classified as not present inside the mask region, and the false negatives (FN) are the pixels that are falsely classified as not present inside the masked region. Thus, we can now calculate the precision and recall from the confusion matrix as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

4.2.2 Dice similarity coefficient

The Dice coefficients, as coined by [15], are used widely for image segmentation purpose, and it has been used in the case of both 2D and 3D image segmentation tasks. The Dice coefficients required for image segmentation can be constructed from a contingency table [88] of four possible outcomes as represented in the probabilities of segmentation results from an image. Dice score can be generalized using the definitions of true positives (TP), false positives (FP), and false negatives (FN) as:

$$\text{DICE} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{6}$$

The Dice coefficient measures how much the area of interest of two images has overlapped. Dice score values

Table 2 Comparisons of the segmentation result for the proposed and conventional methods in all the employed datasets

Dataset	Model	Backbone	Precision	Recall	Dice	mIoU
DRIVE	U-Net [49]	–	94.46	88.32	77.19	62.94
	U-Net++ [85]	–	97.31	90.59	82.24	70.44
	Attention UNet [43]	–	97.55	92.49	80.51	67.52
	MultiResU-Net [29]		97.20	88.12	83.08	71.17
	FA-Net [63]	–	81.89	82.15	81.83	69.27
	ConvUNeXt [24]	ConvNeXt	–	–	82.30	**82.60**
	LCP-Net [44]	–	89.69	78.72	–	82.14
	Res2UNet [33]	–	–	–	81.86	69.26
	LadderNet [87]	–	–	78.56	82.02	–
	DoubleU-Net [30]	VGG-19	94.22	91.63	83.22	70.82
	IterNet++ [86]	–	–	83.99	83.13	71.15
	Sharp U-Net [89]	ResNet-50	–	–	82.03	71.43
	CE-Net [22]	–	-	83.09	81.35	68.54
	R2U-Net [5]	–	–	–	81.24	68.38
	SUD-GAN [78]	–	88.21	83.40	–	–
	DoubleU-NetPlus	EfficientNetB7	**98.05**	**96.48**	**85.17**	73.92
LUNA	U-Net [49]	–	97.42	95.36	95.11	91.30
	U-Net++ [85]	–	99.29	98.48	94.07	93.61
	Attention U-Net [43]	–	99.25	96.96	93.60	93.09
	MultiResU-Net [29]		99.33	97.63	92.14	90.07
	CE-Net [22]	–	–	98.00	–	96.20
	DCU-Net [13]	–	97.31	96.99	97.33	96.17
	DoubleU-Net [30]	VGG-19	95.68	89.20	97.74	95.11
	FBUNet [66]	–	98.33	98.67	98.50	97.06
	EANet [68]	VGG-19	98.66	98.77	98.65	–
	Sharp U-Net [89]	ResNet-50	–	98.77	97.25	95.22
	FA-Net [63]	–	83.16	84.14	95.44	93.70
	Bose et al. [8]	–	–	97.90	–	97.00
	DoubleU-NetPlus	EfficientNetB7	**99.57**	**98.82**	**99.34**	**98.93**
BUSI	U-Net [49]	–	94.04	84.67	80.44	63.86
	U-Net++ [85]	–	93.81	85.19	81.29	65.55
	Attention U-Net [43]	–	91.06	77.31	81.34	68.10
	MultiResU-Net [29]	–	96.77	85.15	78.76	63.60
	FA-Net [63]	–	93.10	88.73	90.63	71.95
	CE-Net [22]	–	–	–	80.69	67.63
	MCRNet [40]	ResNet-34	–	81.92	82.31	69.94
	RCA-IUnet [46]	–	94.01	89.11	91.40	**89.95**
	DoubleU-Net [30]	VGG-19	93.11	88.85	89.93	71.49
	Sharp U-Net [89]	ResNet-50	–	92.42	90.82	74.20
	CFPNet [39]	–	–	–	80.92	67.95
	GC-Net [77]	–	86.5	–	82.10	73.80
	DoubleU-NetPlus	EfficientNetB7	**96.90**	**92.47**	**94.30**	84.71

Table 2 (continued)

Dataset	Model	Backbone	Precision	Recall	Dice	mIoU
CVCclinicDB	U-Net [49]	–	95.21	89.35	88.11	83.52
	U-Net++ [85]	–	92.11	82.23	87.40	77.27
	Attention U-Net [43]	–	94.20	92.00	93.88	88.87
	MultiResU-Net [29]	–	95.96	92.59	94.10	89.96
	DoubleU-Net [30]	VGG-19	95.92	84.57	92.39	86.11
	CRF-EfficientUNet [62]	EfficientNetB7	96.83	**97.94**	95.12	91.85
	MSRF-Net [56]	–	94.27	95.67	94.20	90.43
	Sharp U-Net [89]	ResNet-50	–	–	90.05	83.98
	AMNet [55]	Res2Net	–	–	93.60	88.80
	FA-Net [63]	–	94.01	93.39	93.55	89.37
	Focus U-Net [79]	–	93.00	95.60	94.10	89.30
	Ds-transunet [34]	Transformer	93.69	95.00	93.80	89.10
	Swin U-Net [10]	Transformer	–	–	92.30	87.50
	Polyp-PVT [17]	Transformer	–	–	93.70	88.90
	DoubleU-NetPlus	EfficientNetB7	**97.96**	93.87	**96.40**	**95.12**
2018 DSB	U-Net [49]	ResNet-101	96.84	78.55	92.09	85.60
	U-Net++ [85]	–	91.99	79.41	85.81	85.52
	Attention U-Net [43]	–	98.09	80.07	93.21	87.61
	MultiResU-Net [29]	–	98.45	83.77	87.55	83.51
	DoubleU-Net [30]	VGG-19	94.96	94.07	92.33	88.07
	Sharp U-Net [89]	ResNet-50	–	–	95.40	89.60
	FA-Net [63]	–	91.94	92.22	91.76	85.69
	MSRF-Net [56]	–	90.22	94.02	92.24	85.34
	Poudel and Lee [45]	EfficientNetB3	–	–	90.07	**90.97**
	DoubleU-NetPlus	EfficientNetB7	**98.82**	**95.64**	**95.76**	90.29
ISBI 2012	U-Net [49]	–	99.53	81.58	94.30	89.38
	U-Net++ [85]	-	94.61	95.56	93.94	88.60
	Attention U-Net [43]	–	99.70	82.20	96.80	93.81
	MultiResU-Net [29]	-	99.72	82.50	96.13	92.90
	FBUNet [66]	–	93.73	94.19	93.96	88.62
	LCP-Net [44]	–	89.69	**98.95**	**98.12**	82.14
	DoubleU-Net [30]	VGG-19	95.85	92.36	93.60	89.87
	Sharp U-Net [89]	ResNet50	–	–	93.52	91.21
	FA-Net [63]	–	95.29	95.68	95.47	91.34
	DoubleU-NetPlus	EfficientNetB7	**99.75**	88.62	97.10	**94.38**

The numbers in bold indicate the best performance of the corresponding indicators

"–" denotes that there is no backbone used in the network

have a range of [0,1]. The higher the Dice score is, the better segmentation is achieved from the prediction.

4.2.3 Intersection-Over-Union (IoU)

Along with the Dice score, mean intersection-over-union (mIoU) can be used to calculate the prediction similarity with ground truth. IoU values have a range of [0,1]. The higher value of IoU means there is a better similarity

between prediction and ground truth. IoU can be defined in terms of the common confusion metrics as follows:

$$IoU = \frac{TP}{TP + FP + FN} \tag{7}$$

4.3 Evaluation of the segmentation results

This section provides the quantitative and qualitative analysis result analysis of the proposed DoubleU-NetPlus method with other SOTA methods.

4.3.1 Quantitative result analysis

Here, we report the quantitative results on six various modalities of medical image datasets and compare them to other SOTA approaches to ensure that the proposed model surpasses the performance or performs at par with other SOTA methods (on the same train-test split ratio and similar types of data augmentation methods). It is important to note that in order to provide a fair comparison, the evaluation metrics are provided only for the approaches that prioritize segmentation performance over computational efficiency. The performance of the model on all the utilized datasets is shown in Table 2.

Results on DRIVE A comparison with well-established segmentation architectures with different backbones demonstrates that our proposed method outperforms the SOTA architectures. With a Dice score of 85.17%, mIoU of 73.92%, precision of 98.05%, and recall of 96.48% (see Table 2), the DoubleU-NetPlus architecture significantly surpasses all SOTA architectures on the DRIVE dataset. While outperforming U-Net and most of its variants, it can also be observed that DoubleU-NetPlus exceeds the performance of the recently proposed ConvNeXt [37] encoder backbone-based ConvUNeXt [24] architecture by a Dice score of 2.87% though ConvUNeXt reports the highest mIoU value of 82.60%. Compared to FANet [63], the model achieves an increase of 4.65% in the mIoU metric with a lesser number of augmented images during training.

Results on LUNA In the LUNA dataset, the DouleU-NetPlus network achieves SOTA segmentation results of 99.34% on Dice, 98.93% on mIoU, 99.57% on precision, and 98.82% on recall metrics, respectively (see Table 2). The results outperform U-Net [49], U-Net++ [85], VGG-19 encoder-based EANet [68], and ResNet-50 based Sharp U-Net [89] architectures in the Dice metric by a margin of 4.23%, 5.27%, 0.69%, and 2.09%. DoubleU-NetPlus also has the best balance on both the precision-recall and Dice-mIoU pairs.

Results on BUSI In the BUSI dataset, the DoubleU-NetPlus achieves significantly improved results compared to all the SOTA architectures. It achieves a precision of 96.90% and a recall of 92.47%. The model achieves a significantly improved Dice value of 94.30% which is 13.01% and 15.54% better compared to the UNet++ [85], and MultiResUNet [29] architectures, respectively (see Table 2). Although the highest mIoU is achieved by RCA-IUnet [46] with 89.95%, compared to DoubleU-NetPlus's 84.71%.

Results on CVCclinicDB Table 2 demonstrates that in the CVCclinicDB dataset, DoubleU-NetPlus produces a Dice score of 96.40%, mIoU of 95.12%, precision of 97.96%, and a recall value of 93.87% with an improvement of 4.01% in Dice with respect to SOTA DoubleU-Net architecture. Our model achieves the best trade-off between Dice and mIoU metrics compared with the SOTA architectures resulting in the highest mIoU metric value of 95.12%, surpassing the dual Swin Transformer-based Ds-transunet [34] model by 6.02% in the mIoU metric.

Results on 2018 DSB DoubleU-NetPlus obtains significantly improved precision value of 98.82%, Dice of 95.76%, and mIoU of 90.29% which are much improved results compared to U-Net [49], U-Net++ [85], and DoubleU-Net [30] (see Table 2). It also achieves the best trade-off between Dice-mIoU compared to other SOTA architectures. Though Sharp U-Net [89] reports the highest Dice value of 95.40%, in terms of mIou, DoubleU-NetPlus generates better results. Poudel and Lee [45] report the highest mIoU of 90.97%; however, DoubleU-NetPlus outperforms their architecture by 5.69% in the Dice metric.

Results on ISBI 2012 In the ISBI 2012 dataset, DoubleU-NetPlus achieves 99.75% in precision, 88.62% in the recall, 97.10% in Dice, and 94.38% in mIoU metric, which are significantly improved results compared to the U-Net [49], U-Net++ [85], and MultiResU-Net [29] architectures. Especially in the mIoU metric, the proposed model obtains an increase of 5.00%, 5.78%, 0.57%, and 1.48% compared to U-Net, U-Net++, Attention U-Net [43], and MultiResU-Net architectures respectively (see Table 2). The highest Dice value of 98.12% is reported in LCP-Net [44].

The results of the DoubleU-NetPlus model show that the proposed model greatly improves the performance of MIS tasks in diverse modalities of colonoscopy, fluorescence, electron microscopy, CT, retinal, and ultrasound.

4.3.2 Qualitative result analysis

The results that were obtained from the experiments on six datasets of diverse modalities were evaluated critically on visual qualitative criteria to ensure proper segmentation performance. Specifically, we illustrate the predictions of U-Net, U-Net++, Attention U-Net, MultiResU-Net, and our proposed DoubleU-NetPlus segmentation architectures, which were also applied in the quantitative comparisons too. The visual comparisons of the mentioned architectures with the proposed DoubleU-NetPlus, as demonstrated in Figs. 8a, b, c, and 9a, b, c, shows that the segmentation map of the DoubleU-NetPlus network achieves better semantic segmentation performance in every datasets. On

Table 3 Ablation experiments that analyze the contributions of the different modules on the utilized datasets

Dataset	Model	Precision (±0.5%)	Recall (±0.5%)	Dice (±0.5%)	mIoU (±0.5%)
Drive	Baseline	98.42	96.36	97.90	96.64
	DoubleU-NetPlus w/o MKRC	97.51	95.81	83.27	71.53
	DoubleU-NetPlus w/o TAM	97.53	95.67	83.65	72.12
	DoubleU-NetPlus w/o TAG	97.55	96.98	83.71	72.09
	DoubleU-NetPlus w/o (TAM & MKRC)	96.52	97.08	82.29	70.00
	DoubleU-NetPlus w/o (TAM & MKRC & TAG)	97.19	96.11	82.30	70.24
	DoubleU-NetPlus	**98.05**	**96.48**	**85.17**	73.92
LUNA	Baseline	98.42	96.36	97.90	96.64
	DoubleU-NetPlus w/o MKRC	98.04	98.10	98.40	97.07
	DoubleU-NetPlus w/o TAM	98.79	98.19	98.61	97.53
	DoubleU-NetPlus w/o TAG	98.83	97.29	98.10	96.90
	DoubleU-NetPlus w/o (TAM & MKRC)	98.10	98.25	98.11	96.83
	DoubleU-NetPlus w/o (TAM & MKRC & TAG)	97.91	98.05	97.92	96.77
	DoubleU-NetPlus	**99.57**	**98.82**	**99.34**	**98.93**
BUSI	Baseline	94.04	84.67	88.47	80.67
	DoubleU-NetPlus w/o MKRC	96.11	90.48	93.74	79.32
	DoubleU-NetPlus w/o TAM	**97.36**	91.78	93.84	75.11
	DoubleU-NetPlus w/o TAG	95.88	91.71	93.13	82.71
	DoubleU-NetPlus w/o (TAM & MKRC)	96.46	91.81	94.01	84.45
	DoubleU-NetPlus w/o (TAM & MKRC & TAG)	96.93	92.17	94.20	83.58
	DoubleU-NetPlus	96.90	**92.47**	**94.30**	**84.71**
CVCclinicDB	Baseline	97.36	91.85	95.07	90.87
	DoubleU-NetPlus w/o MKRC	98.92	95.72	97.83	94.51
	DoubleU-NetPlus w/o TAM	98.89	**96.12**	**98.05**	**95.52**
	DoubleU-NetPlus w/o TAG	**98.96**	95.72	97.83	94.51
	DoubleU-NetPlus w/o (TAM & MKRC)	98.89	95.68	97.78	93.27
	DoubleU-NetPlus w/o (TAM & MKRC & TAG)	99.02	95.83	97.97	96.08
	DoubleU-NetPlus	97.96	93.87	96.40	95.12
2018 DSB	Baseline	96.84	78.55	92.09	85.60
	DoubleU-NetPlus w/o MKRC	96.39	83.10	91.32	83.91
	DoubleU-NetPlus w/o TAM	**99.03**	82.67	90.84	**90.57**
	DoubleU-NetPlus w/o TAG	97.58	83.29	89.21	85.73
	DoubleU-NetPlus w/o (TAM & MKRC)	96.46	80.63	88.57	82.18
	DoubleU-NetPlus w/o (TAM & MKRC & TAG)	94.62	90.29	92.48	86.88
	DoubleU-NetPlus	98.82	**95.64**	**95.76**	90.29
ISBI 2012	Baseline	99.53	81.58	94.30	89.38
	DoubleU-NetPlus w/o MKRC	99.73	82.35	96.99	94.17
	DoubleU-NetPlus w/o TAM	99.61	82.31	96.72	93.66
	DoubleU-NetPlus w/o TAG	99.69	82.12	96.79	93.79
	DoubleU-NetPlus w/o (TAM & MKRC)	99.69	82.04	96.66	93.54
	DoubleU-NetPlus w/o (TAM & MKRC & TAG)	99.61	82.27	96.62	93.47
	DoubleU-NetPlus	**99.75**	**88.62**	**97.10**	**94.38**

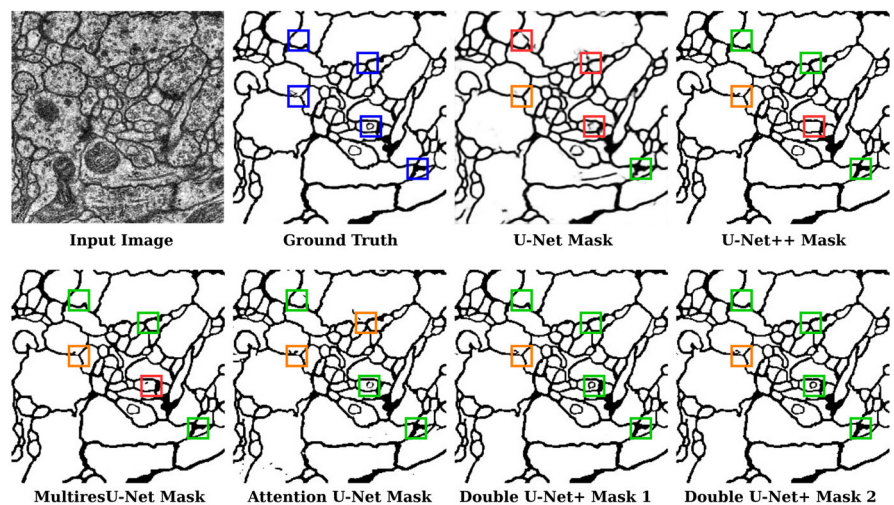The numbers in bold indicate the best performance of the corresponding indicators

Fig. 8 Visual comparative analysis among different segmentation methods. First row (left to right): input image, ground truth, U-Net output, U-Net++ output. Second row (left to right): MultiResU-Net output, Attention U-Net output, and results of the DoubleU-NetPlus network (Mask 1 and 2) on the LUNA dataset. A similar pattern is followed in rows three to six for CVCclinicDB and ISBI 2012 datasets. Blue, red, yellow, and green boxes denote exemplary ROI, unsatisfactory, moderate, and good results
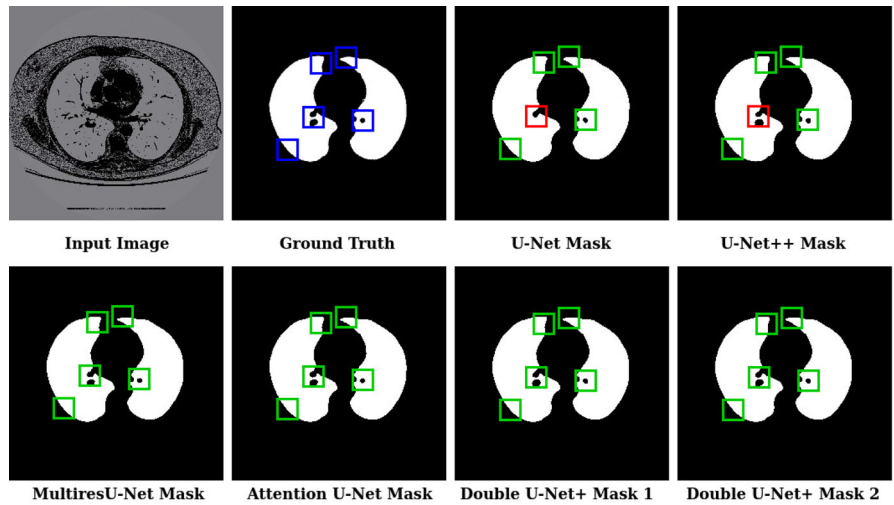
(a) LUNA
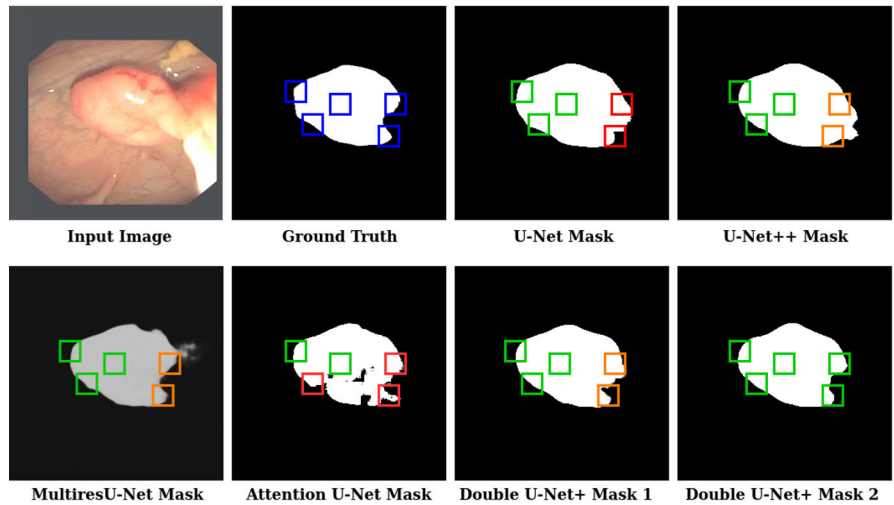
(b) CVCclinicDB

(c) ISBI 2012

Fig. 9 Visual comparative analysis among different segmentation methods. First row (left to right): input image, ground truth, U-Net output, U-Net++ output. Second row (left to right): MultiResU-Net output, Attention U-Net output, and results of the DoubleU-NetPlus network (Mask 1 and 2) on the DRIVE dataset. A similar pattern is followed in rows three to six for BUSI and 2018 DSB datasets. Blue, red, yellow, and green boxes denote exemplary ROI, unsatisfactory, moderate, and good results
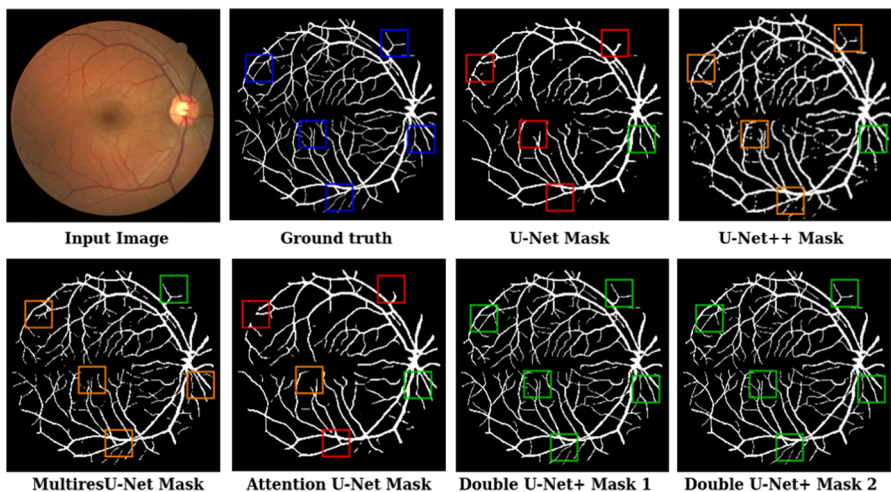
(a) DRIVE
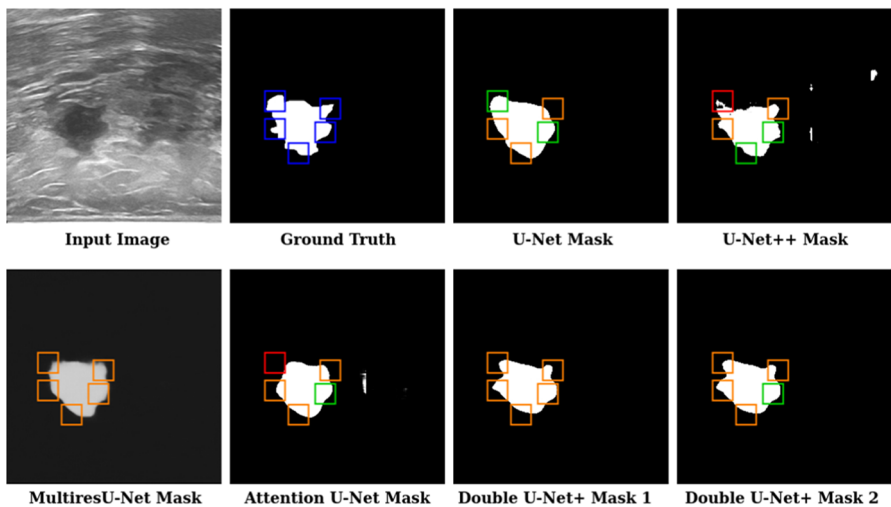
(b) BUSI
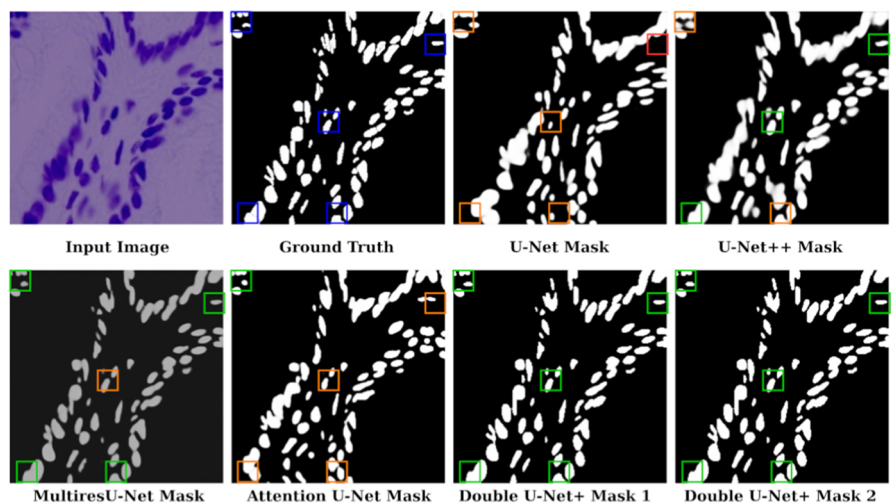
(c) 2018 DSB

Table 4 *P*-values between proposed DoubleU-NetPlus and other SOTA methods on different evaluation metrics

Pair	Dice	mIoU
U-Net versus DoubleU-NetPlus	0.0101	0.0091
U-Net++ versus DoubleU-NetPlus	0.0074	0.0231
Att.U-Net versus DoubleU-NetPlus	0.0455	0.0364
MultiResU-Net versus DoubleU-NetPlus	0.0432	0.0451
FANet versus DoubleU-NetPlus	0.0003	0.0078
DoubleU-Net versus DoubleU-NetPlus	0.0010	0.0185
Sharp U-Net versus DoubleU-NetPlus	0.0114	0.0326

visual inspection, it is clear that there are several instances where the proposed network outperforms SOTA architectures such as the U-Net, U-Net++, Attention U-Net, and MultiResU-Net (Table 3).

4.3.3 Statistical significance test

To statistically investigate the performance of the proposed DoubleU-NetPlus over other SOTA segmentation methods on different quantitative metrics, we conduct paired sample t tests between the Dice and mIoU obtained by DoubleU-NetPlus and the Dice and mIoU obtained by other methods. A paired sample t test is often used for comparing two methods on the same evaluation metric in the MIS domain [56, 65, 69, 77]. We perform the test on Dice and mIoU metrics mainly because these two are the most significant evaluation metrics in semantic image segmentation. It should be noted that we do not include the precision and recall metrics in the test because every compared method does not report these two metrics. A comparison was done with those methods which utilized all six datasets in their study or reported in the literature. A *p*-value less than 0.05 is considered as statistically significant, and the paired-wise *p*-values are reported in Table 4. From Table 4, it is clear that in all seven paired methods, the *p*-values are smaller than 0.05 for the Dice and mIoU metrics, which demonstrates that our proposed method achieved significantly improved results compared to seven other SOTA models.

4.3.4 Ablation studies

We have performed an extensive ablation study in each of the employed datasets to empirically verify some of our incorporated modules in the proposed DoubleU-NetPlus network. A baseline U-Net was used to benchmark the performance of various datasets that were used in our experiments. We investigate the baseline performance of the U-Net by training it with the same number of augmented images that were used to train the proposed DoubleU-NetPlus model and sequentially assess the performance with subsequent removal of the individual MKRC, TAM, and TAG modules. We also investigated by removing (TAM, MKRC), and (TAM, MKRC, and TAG) modules combined from the proposed architecture. The results of module removal on the BUSI and DRIVE dataset are demonstrated in Table 3. It can be observed that the EfficientNetB7-based encoder backbone, TAM, MKRC, and TAG modules contribute significantly to the improvements in Dice score, mIoU, precision, and recall metric values.

5 Conclusion

Semantic segmentation of medical images is a key element in medical image analysis. This paper presents a robust deep learning-based MIS network named DoubleU-Net-Plus equipped with several architectural modifications, mainly the integration of pre-trained EfficientNetB7 as a feature encoder backbone, a newly proposed multi-kernel residual convolution module, multi-scale feature re-calibrating SE-ASPP module, and a hybrid triple attention module at the bottleneck of each network. We also integrated attention-driven residual convolutions throughout the encoder and decoder part of the network. To capture salient regions with higher precision, we have integrated a novel triple attention gate module that focuses on the relevant regions and suppresses other irrelevant regions in the skip connections features. A combination of all these modules together captures high-level semantic and discriminative feature maps while preserving effective spatial

information. Experimental results evaluated on six benchmark datasets of different modalities demonstrate the proposed model's superiority over SOTA segmentation methods in MIS tasks. We believe that DoubleU-NetPlus is a generic segmentation model and can be applied to similar 2D MIS tasks. One of the challenges in this architecture is its high number of trainable parameters. We plan to reduce the number of parameters and computational complexity in the future. We also plan to adjust the design of the network to make it adaptable in the 3D image domain.

Author contributions MRA involved in conceptualization of this study, methodology, formal analysis, software, validation, writing—original draft preparation, writing—review, editing, and investigation. AFA involved in software, visualization, writing—original draft preparation, review, and editing. RUA involved in software, validation, and writing—original draft preparation. SS involved in review, validation, and editing. AKMMI involved in review and investigation. SI involved in review and investigation.

Data availability The research has used publicly available datasets.

Declarations

Conflict of 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.

References

1. Ahmed MR, Robin TI, Shafin AA (2020) Automatic environmental sound recognition (aesr) using convolutional neural network. Int J Modern Educ Comput Sci 12(5):41–54
2. Ahmed MR, Islam S, Islam AM et al (2023) An ensemble 1d-cnn-lstm-gru model with data augmentation for speech emotion recognition. Expert Syst Appl 218(119):633. https://doi.org/10.1016/j.eswa.2023.119633
3. Al-Dhabyani W, Gomaa M, Khaled H et al (2020) Dataset of breast ultrasound images. Data Brief 28(104):863. https://doi.org/10.1016/j.dib.2019.104863
4. Al-Masni MA, Kim DH (2021) Cmm-net: contextual multi-scale multi-level network for efficient biomedical image segmentation. Sci Rep 11(1):1–18
5. Alom MZ, Hasan M, Yakopcic C, et al (2018) Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. arXiv preprint arXiv:1802.06955
6. Badrinarayanan V, Kendall A, Cipolla R (2017) Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell 39(12):2481–2495
7. Bernal J, Sánchez FJ, Fernández-Esparrach G et al (2015) Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Comput Med Imaging Graph 43:99–111
8. Bose S, Chowdhury RS, Das R et al (2022) Dense dilated deep multiscale supervised u-network for biomedical image segmentation. Comput Biol Med 143(105):274
9. Caicedo JC, Goodman A, Karhohs KW et al (2019) Nucleus segmentation across imaging experiments: the 2018 data science bowl. Nat Methods 16(12):1247–1253
10. Cao H, Wang Y, Chen J, et al (2021) Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537
11. Cardona A, Saalfeld S, Preibisch S et al (2010) An integrated micro- and macroarchitectural analysis of the drosophila brain by computer-assisted serial section electron microscopy. PLoS Biol 8(10):1–17. https://doi.org/10.1371/journal.pbio.1000502
12. Cardona A, Saalfeld S, Schindelin J et al (2012) Trakem2 software for neural circuit reconstruction. PLoS ONE 7(6):1–8. https://doi.org/10.1371/journal.pone.0038011
13. Kb Chen, Xuan Y, Aj Lin et al (2021) Lung computed tomography image segmentation based on u-net network fused with dilated convolution. Comput Methods Programs Biomed 207(106):170
14. Chen LC, Papandreou G, Kokkinos I et al (2017) Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans Pattern Anal Mach Intell 40(4):834–848
15. Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26(3):297–302
16. Ding Y, Chen F, Zhao Y et al (2019) A stacked multi-connection simple reducing net for brain tumor segmentation. IEEE Access 7:104,011-104,024
17. Dong B, Wang W, Fan DP, et al (2021) Polyp-pvt: Polyp segmentation with pyramid vision transformers. arXiv preprint arXiv:2108.06932
18. Gao C, Ye H, Cao F et al (2021) Multiscale fused network with additive channel-spatial attention for image segmentation. Knowl-Based Syst 214(106):754
19. Gao SH, Cheng MM, Zhao K et al (2019) Res2net: A new multi-scale backbone architecture. IEEE Trans Pattern Anal Mach Intell 43(2):652–662
20. Gao Z, Xie J, Wang Q, et al (2019b) Global second-order pooling convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 3024–3033
21. Ghosh A, Ehrlich M, Shah S, et al (2018) Stacked u-nets for ground material segmentation in remote sensing imagery. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 257–261
22. Gu Z, Cheng J, Fu H et al (2019) Ce-net: context encoder network for 2d medical image segmentation. IEEE Trans Med Imaging 38(10):2281–2292
23. Guo MH, Xu TX, Liu JJ et al (2022) Attention mechanisms in computer vision: a survey. Comput Vis Media. 1–38
24. Han Z, Jian M, Wang GG (2022) Convunext: an efficient convolution neural network for medical image segmentation. Knowl-Based Syst 253(109):512
25. Hatamizadeh A, Tang Y, Nath V, et al (2022) Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 574–584
26. He K, Zhang X, Ren S, et al (2016a) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
27. He K, Zhang X, Ren S, et al (2016b) Identity mappings in deep residual networks. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14, Springer, pp 630–645
28. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7132–7141

29. Ibtehaz N, Rahman MS (2020) Multiresunet: rethinking the u-net architecture for multimodal biomedical image segmentation. Neural Netw 121:74–87

30. Jha D, Riegler MA, Johansen D, et al (2020) Doubleu-net: A deep convolutional neural network for medical image segmentation. In: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), IEEE, pp 558–564

31. Kaul A, Raina S (2022) Support vector machine versus convolutional neural network for hyperspectral image classification: a systematic review. Concurr Comput Pract Exp 34(15):e6945

32. Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. Commun ACM 60(6):84–90

33. Li X, Ding J, Tang J et al (2022) Res2unet: a multi-scale channel attention network for retinal vessel segmentation. Neural Comput Appl 34(14):12,001-12,015

34. Lin A, Chen B, Xu J, et al (2022) Ds-transunet: dual swin transformer u-net for medical image segmentation. IEEE Trans Instrum Meas

35. Lin G, Milan A, Shen C, et al (2017) Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1925–1934

36. Liu L, Cheng J, Quan Q et al (2020) A survey on u-shaped networks in medical image segmentations. Neurocomputing 409:244–258

37. Liu Z, Mao H, Wu CY, et al (2022) A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 11,976–11,986

38. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3431–3440

39. Lou A, Guan S, Loew M (2021) Cfpnet-m: A light-weight encoder-decoder based network for multimodal biomedical image real-time segmentation. arXiv preprint arXiv:2105.04075

40. Lou M, Meng J, Qi Y et al (2022) Mcrnet: multi-level context refinement network for semantic segmentation in breast ultrasound imaging. Neurocomputing 470:154–169

41. Mader KS (2016) Lung nodule analysis (luna) - lung segmentation dataset. https://www.kaggle.com/datasets/kmader/finding-lungs-in-ct-data, [Accessed on: 29/03/2022]

42. Nishio M, Fujimoto K, Togashi K (2021) Lung segmentation on chest x-ray images in patients with severe abnormal findings using deep learning. Int J Imaging Syst Technol 31(2):1002–1008

43. Oktay O, Schlemper J, Folgoc LL, et al (2018) Attention u-net: learning where to look for the pancreas. arXiv preprint arXiv:1804.03999

44. Peng D, Xiong S, Peng W et al (2021) Lcp-net: a local context-perception deep neural network for medical image segmentation. Expert Syst Appl 168(114):234

45. Poudel S, Lee SW (2021) Deep multi-scale attentional features for medical image segmentation. Appl Soft Comput 109(107):445

46. Punn NS, Agarwal S (2022) Rca-iunet: a residual cross-spatial attention-guided inception u-net model for tumor segmentation in breast ultrasound imaging. Mach Vis Appl 33(2):1–10

47. Qin X, Zhang Z, Huang C et al (2020) U2-net: going deeper with nested u-structure for salient object detection. Pattern Recogn 106(107):404

48. Rahman MF, Zhuang Y, Tseng TLB et al (2022) Improving lung region segmentation accuracy in chest x-ray images using a two-model deep learning ensemble approach. J Vis Commun Image Represent 85(103):521

49. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, pp 234–241

50. Sang DV, Chung TQ, Lan PN, et al (2021) Ag-curesnest: a novel method for colon polyp segmentation. arXiv preprint arXiv:2105.00402

51. Shopon M, Hsu GSJ, Gavrilova ML (2022) Multiview gait recognition on unconstrained path using graph convolutional neural network. IEEE Access 10:54,572-54,588

52. Shuvo MB, Ahommed R, Reza S et al (2021) Cnl-unet: a novel lightweight deep learning architecture for multimodal biomedical image segmentation with false output suppression. Biomed Signal Process Control 70(102):959

53. Siddique N, Paheding S, Alom MZ, et al (2021) Recurrent residual u-net with efficientnet encoder for medical image segmentation. In: Pattern Recognition and Tracking XXXII, SPIE, pp 134–142

54. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556

55. Song P, Li J, Fan H (2022) Attention based multi-scale parallel network for polyp segmentation. Comput Biol Med 146(105):476

56. Srivastava A, Jha D, Chanda S et al (2021) Msrf-net: a multi-scale residual fusion network for biomedical image segmentation. IEEE J Biomed Health Inform 26(5):2252–2263

57. Staal J, Abràmoff MD, Niemeijer M et al (2004) Ridge-based vessel segmentation in color images of the retina. IEEE Trans Med Imaging 23(4):501–509

58. Su R, Zhang D, Liu J et al (2021) Msu-net: multi-scale u-net for 2d medical image segmentation. Front Genet 12(639):930

59. Sudre CH, Li W, Vercauteren T, et al (2017) Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 10553 LNCS:240–248. https://doi.org/10.1007/978-3-319-67558-9_28

60. Szegedy C, Liu W, Jia Y, et al (2015) Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1–9 https://doi.org/10.1109/CVPR.2015.7298594

61. Tan M, Le Q (2019) Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, PMLR, pp 6105–6114

62. Thanh NC, Long TQ et al (2021) Crf-efficientunet: an improved unet framework for polyp segmentation in colonoscopy images with combined asymmetric loss function and crf-rnn layer. IEEE Access 9:156,987-157,001

63. Tomar NK, Jha D, Riegler MA, et al (2022) Fanet: a feedback attention network for improved biomedical image segmentation. IEEE Trans Neural Netwo Learn Syst

64. Tong X, Wei J, Sun B et al (2021) Ascu-net: attention gate, spatial and channel attention u-net for skin lesion segmentation. Diagnostics 11(3):501

65. Valanarasu JMJ, Sindagi VA, Hacihaliloglu I et al (2021) Kiu-net: overcomplete convolutional architectures for biomedical image and volumetric segmentation. IEEE Trans Med Imaging 41(4):965–976

66. Wang H, Yang J (2021) Fbunet: full convolutional network based on fusion block architecture for biomedical image segmentation. J Med Biol Eng 41(2):185–202

67. Wang J, Lv P, Wang H et al (2021) Sar-u-net: squeeze-and-excitation block and atrous spatial pyramid pooling based residual u-net for automatic liver segmentation in computed tomography. Comput Methods Programs Biomed 208(106):268

68. Wang K, Zhang X, Zhang X et al (2022) Eanet: iterative edge attention network for medical image segmentation. Pattern Recogn 127(108):636

69. Wang L, Gu J, Chen Y et al (2021) Automated segmentation of the optic disc from fundus images using an asymmetric deep learning network. Pattern Recogn 112(107):810

70. Wang X, Jiang X, Ding H et al (2021) Knowledge-aware deep framework for collaborative skin lesion segmentation and melanoma recognition. Pattern Recogn 120(108):075

71. Woo S, Park J, Lee JY, et al (2018) Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 3–19

72. Wu H, Wang W, Zhong J et al (2021) Scs-net: a scale and context sensitive network for retinal vessel segmentation. Med Image Anal 70(102):025. https://doi.org/10.1016/j.media.2021.102025

73. Wu Z, Su L, Huang Q (2019) Cascaded partial decoder for fast and accurate salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 3907–3916

74. Xia X, Kulis B (2017) W-net: A deep model for fully unsupervised image segmentation. arXiv preprint arXiv:1711.08506

75. Xiao C, Hong B, Liu J et al (2022) Deep residual contextual and subpixel convolution network for automated neuronal structure segmentation in micro-connectomics. Comput Methods Programs Biomed. https://doi.org/10.1016/j.cmpb.2022.106759

76. Xie X, Pan X, Zhang W et al (2022) A context hierarchical integrated network for medical image segmentation. Comput Electr Eng 101(108):029. https://doi.org/10.1016/j.compeleceng.2022.108029

77. Xue C, Zhu L, Fu H et al (2021) Global guidance network for breast lesion segmentation in ultrasound images. Med Image Anal 70(101):989

78. Yang T, Wu T, Li L et al (2020) Sud-gan: deep convolution generative adversarial network combined with short connection and dense block for retinal vessel segmentation. J Digit Imaging 33(4):946–957

79. Yeung M, Sala E, Schönlieb CB et al (2021) Focus u-net: a novel dual attention-gated cnn for polyp segmentation during colonoscopy. Comput Biol Med 137(104):815

80. Zeng Z, Xie W, Zhang Y et al (2019) Ric-unet: an improved neural network based on unet for nuclei segmentation in histology images. Ieee Access 7:21,420-21,428

81. Zhang J, Jin Y, Xu J, et al (2018) Mdu-net: Multi-scale densely connected u-net for biomedical image segmentation. 10.48550/arxiv.1812.00352

82. Zhang Y, Dong J (2022) 2k-fold-net and feature enhanced 4-fold-net for medical image segmentation. Pattern Recogn 127(108):625

83. Zhang Y, He M, Chen Z et al (2022) Bridge-net: context-involved u-net with patch-based loss weight mapping for retinal blood vessel segmentation. Expert Syst Appl 195(116):526

84. Zhao P, Zhang J, Fang W et al (2020) Scau-net: spatial-channel attention u-net for gland segmentation. Front Bioeng Biotech 8:670

85. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, et al (2018) Unet++: a nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4, Springer, pp 3–11

86. Zhu M, Zeng K, Lin G et al (2022) Iternet++: an improved model for retinal image segmentation by curvelet enhancing, guided filtering, offline hard-sample mining, and test-time augmenting. IET Image Proc 16(13):3617–3633

87. Zhuang J (2018) Laddernet: Multi-path networks based on u-net for medical image segmentation. arXiv preprint arXiv:1810.07810

88. Zou KH, Warfield SK, Bharatha A et al (2004) Statistical validation of image segmentation quality based on a spatial overlap index1. Acad Radiol 11:178–189. https://doi.org/10.1016/S1076-6332(03)00671-8

89. Zunair H, Hamza AB (2021) Sharp u-net: depthwise convolutional network for biomedical image segmentation. Comput Biol Med 136(104):699

Authors and Affiliations

Md. Rayhan Ahmed[1] (ORCID) · **Adnan Ferdous Ashrafi**[2] · **Raihan Uddin Ahmed**[3] · **Swakkhar Shatabda**[1] · **A. K. M. Muzahidul Islam**[1] · **Salekul Islam**[1]

✉ Md. Rayhan Ahmed
rayhan@cse.uiu.ac.bd

Adnan Ferdous Ashrafi
adnan@stamforduniversity.edu.bd

Raihan Uddin Ahmed
raihanahmed95@stamforduniversity.edu.bd

Swakkhar Shatabda
swakkhar@cse.uiu.ac.bd

A. K. M. Muzahidul Islam
muzahid@cse.uiu.ac.bd

Salekul Islam
salekul@cse.uiu.ac.bd

[1] Department of Computer Science and Engineering, United International University, United City, Madani Avenue, Badda, Dhaka 1212, Bangladesh

[2] Department of Computer Science and Engineering, Stamford University Bangladesh, 51, Siddeswari Road, Ramna, Dhaka 1217, Bangladesh

[3] Department of Electrical and Electronics Engineering, Stamford University Bangladesh, 51, Siddeswari Road, Ramna, Dhaka 1217, Bangladesh