



Ψ net: a parallel network with deeply coupled spatial and squeezed features for segmentation of medical images

Eman M. Elmeslimany¹ · Sherif S. Kishk¹ · Doaa A. Altantawy¹

Received: 10 November 2022 / Revised: 9 May 2023 / Accepted: 24 July 2023 /

Published online: 8 August 2023

© The Author(s) 2023

Abstract

The process of delineating a region of interest or an object in an image is called image segmentation. Efficient medical image segmentation can contribute to the early diagnosis of illnesses, and accordingly, patient survival possibilities can be enhanced. Recently, deep semantic segmentation methods demonstrate state-of-the-art (SOTA) performance. In this paper, we propose a generic novel deep medical segmentation framework, denoted as Ψ net. This model introduces a novel parallel encoder-decoder structure that draws up the power of triple U-Nets. In addition, a multi-stage squeezed-based encoder is employed to raise the network sensitivity to relevant features and suppress the unnecessary ones. Moreover, atrous spatial pyramid pooling (ASPP) is employed in the bottleneck of the network which helps in gathering more effective features during the training process, hence better performance can be achieved in segmentation tasks. We have evaluated the proposed Ψ net on a variety of challengeable segmentation tasks, including colonoscopy, microscopy, and dermoscopy images. The employed datasets include Data Science Bowl (DSB) 2018 challenge as a cell nuclei segmentation from microscopy images, International Skin Imaging Collaboration (ISIC) 2017 and 2018 as skin lesion segmentation from dermoscopy images, Kvansir-SEG, CVC-ClinicDB, ETIS-LaribDB, and CVC-ColonDB as polyp segmentation from colonoscopy images. Despite the variety in the employed datasets, the proposed model, with extensive experiments, demonstrates superior performance to advanced SOTA models, such as U-Net, ResUNet, Recurrent Residual U-Net, ResUNet++, UNet++, BCDU-Net, MultiResUNet, MCGU-Net, FRCU-Net, Attention Deeplabv3p, DDA-Net, ColonSeg-Net, and TMD-Unet.

Keywords Medical image segmentation · U-Net · Ψ net · Polyp segmentation · DSB 2018 · ISIC 2017–2018

✉ Doaa A. Altantawy
doaa.aladel@mans.edu.eg

Eman M. Elmeslimany
eman.mahmoud@std.mans.edu.eg

Sherif S. Kishk
shkishk@mans.edu.eg

¹ Electronics and Communications Engineering Department, Faculty of Engineering, Mansoura University, 60 El-Gomhoria Street, Mansoura, Egypt

1 Introduction

In clinical diagnosis, the segmentation of medical images is a critical and essential procedure for later medical image analysis. Hence, several automatic segmentation techniques have been proposed to help radiologists with early manifestations of life-threatening diseases [26, 46]. These automatic segmentation techniques are roughly categorized into learning-based techniques [41, 53] and classical image processing-based techniques [48, 61]. However, lesions in medical images can vary in size, shape, location, color, texture, and contrast. Hence, the development of accurate and robust segmentation solutions is still a very challenging problem due to several complexities. On the other side, the conventional manual annotation of medical images is a costly time-consuming procedure. Moreover, there is a shortage of specific annotation protocols that suit different types of imaging modalities. Furthermore, low-quality images can potentially influence annotation quality. As a result, employing a computer-aided segmentation model can be an alternate efficient solution to manual image segmentation.

With their remarkable outstanding feature representation capability, convolutional networks have revolutionized different fields, including the computer vision field [5, 47], the industrial field [28], and the monitoring field [27, 29]. Recently, segmentation algorithms based on convolutional neural networks (CNNs) have demonstrated SOTA performance for automated biomedical image segmentation [23, 45, 62]. Most of these algorithms have been encoder-decoder-based networks which have shown prominence for many medical segmentation tasks [40, 43].

Deep encoder-decoder-based CNN has demonstrated high segmentation efficiency due to its skip connections, which permit semantic dense feature maps to propagate from the encoder network to the decoder sub-networks. FCN [42] is one of the earliest deep networks proposed for semantic segmentation that is trained end-to-end for pixel-wise prediction. In [42], the authors have proved that FCNs can significantly enhance accuracy by transferring pre-trained classifier weights, fusing various layers, and learning end-to-end, and pixels-to-pixels on whole images. For the process of transferring weights, they adopted contemporary classification networks, i.e., Alex-Net, VGG-Net, and Google-Net, and transferred their learned representations via fine-tuning to the segmentation task. Then, to produce detailed and accurate segmentations, they developed a skip architecture that combined semantic information from a deep coarse layer with appearance information from a shallow fine layer.

Later, FCN [42] was extended to the most common segmentation network, i.e., U-Net. U-Net [49] is a pixel-wise encoder-decoder architecture that has been trained in an end-to-end way. It has achieved good segmentation performance. It is commonly used for lesion segmentation, anatomical segmentation, and classification in the medical image analysis sector. The main benefit of the U-Net network is that it cannot only precisely segment the targeted object and objectively process and analyze medical images, but it also can aid to improve the accuracy of medical image diagnosis. In addition, with a few training samples, U-Net can perform effectively while still capable of employing global location and context information simultaneously. Moreover, U-Net architecture outperforms FCN in different challengeable segmentation tasks and gradually becomes the pioneering model in the field of medical image segmentation. However, due to the existence of many layers in the conventional U-Net version, a significant amount of time is needed for training. In addition, relatively high GPU memory for larger images is a necessity. Moreover, employing skip connections is a double-edged sword. Skipping

allows a fewer layer-based network which reduces the complexity. In addition, it decays the influence of vanishing gradients effectively. Moreover, a speedy learning process can be achieved. However, on the other side, a semantic gap between low-and high-level features could occur, and some features may be lost between skip connections. Hence, different amendments have been made to the original U-Net architecture to support its weaknesses and add to its strengths [6, 7, 9, 14, 32, 34, 35, 56, 63, 65, 67–69].

In this paper, the main contributions can be summed up as follows.

- 1 We establish an effective novel framework for medical image segmentation, dubbed as Ψ net, which is a squeezed parallel multi-stage encoder-decoder network. Figure 1 indicates a summarized graphical abstract of the proposed segmentation network.
- 2 Due to the adopted parallel mechanism, the atrous spatial pyramid pooling (ASPP), and the squeeze-and-excitation behavior in the introduced encoder, semantically significant features are extracted to enhance segmentation performance. The adopted squeeze-and-excitation module boosts the weights of the most essential features. In addition, it improves the representational power of the proposed segmentation network by enabling dynamic channel-wise feature recalibration. Moreover, the parallel scheme helps to draw up the power of triple U-Nets.
- 3 In practice, multi-scale feature extraction is computationally costly and demands a lot of training data, that is not usually available. However, the proposed parallel Triple scheme with multi-stage encoder-decoder U-Net architecture extracts significant essential features which improve the training efficiency. The proposed segmentation model provides a lightweight and less complex network with a total number of parameters of around 33 M compared to the FRCU-Net of 68 M [9]. The larger the number of parameters, the longer the time needed for convergence.
- 4 With the multi-scale feature extraction property of the proposed Ψ net, efficient segmentation results of small datasets, like ColonDB and ETIS-Larib obtained in terms of dice coefficient and Jaccard index, while the traditional U-Net can't perform effectively with small datasets.
- 5 To demonstrate the generalizability of our model, we have evaluated the proposed Ψ net via a variety of medical image segmentation tasks, such as Kvasir-SEG [36], CVC-ClinicDB [11], CVC-ColonDB [54], ETIS-Larib [50], 2018 Data Science Bowl (DSB) [12], ISIC-2017 [18], and ISIC-2018 [57]. Superior performance is achieved compared to most SOTA models. Figure 2 indicates some visual results of the proposed Ψ net on the employed challenging datasets.

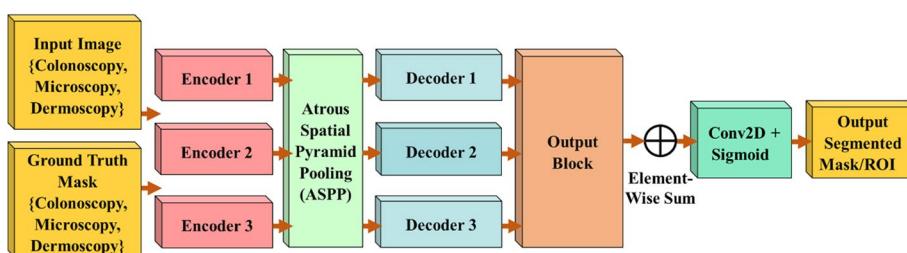


Fig. 1 Overall view of the proposed Ψ net architecture

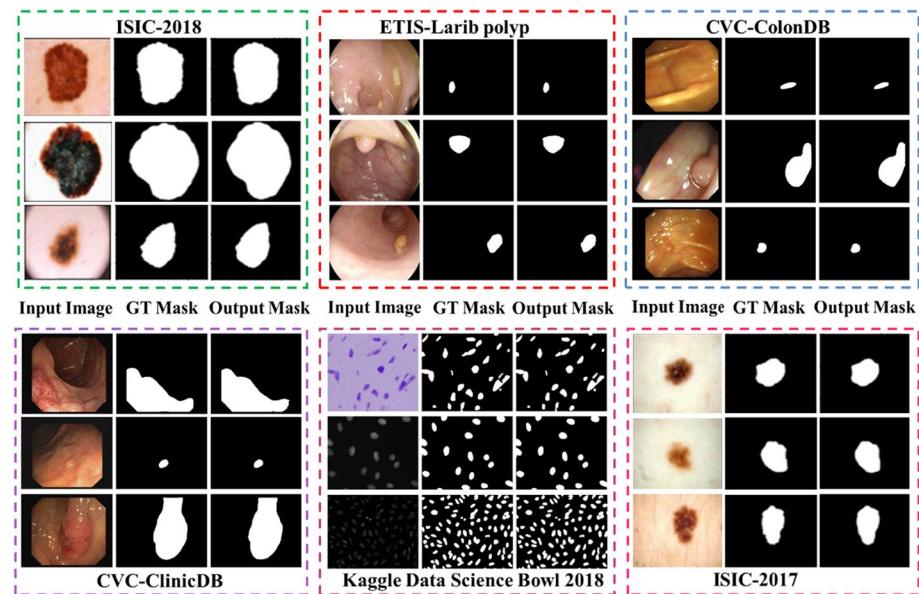


Fig. 2 Visual results of the proposed Ψ net on the employed datasets

- 6 In our experiments, we performed two main types of evaluations. The first one is traditional testing where the training and testing procedures are performed on the same single dataset. The other one is cross-testing in which the model is trained with a specific dataset and tested with another one. In both evaluations, the proposed model shows effectiveness compared to most of the SOTA models.

Our paper is organized in the following manner. Section 2 provides an overview of relevant work in medical image segmentation. The proposed Ψ net architecture is presented in detail in Section 3. In the same section, a background of the traditional U-Net is introduced. In Section 4, the employed datasets and metrics are indicated. In addition, the performed experiments are described, and their results are discussed quantitatively and qualitatively. Finally, we conclude our work in Section 5.

2 Related work

Traditional handcrafted features have been used in semantic segmentation before the uprising of deep learning in computer vision. In the last few years, a variety of deep learning-based approaches have developed rapidly and have achieved outstanding results in image segmentation. The main hurdle of deep architectures is their severe hunger for labeled training data. In addition, due to the limitations of manual annotation, providing large-annotated datasets in medical image segmentation is a challenging task [58].

A convolutional neural network (CNN or ConvNet) is a class of artificial neural network (ANN), which is mostly used to analyze visual images [4, 7, 9, 14, 27–29, 34, 49, 69].

There are several problems with employing such CNNs, such as losing the image spatial information when the convolutional features are fed into fully connected (FC) layers. In addition, training a CNN-based model may include different problems, such as exploding gradient, overfitting, and class imbalance. These challenges can diminish the model's performance. To overcome these problems, FCN architecture was proposed in [42].

Ronneberger et al. [49] modified the conventional FCN, proposed by Long et al. [42], by propagating contextual information from the encoder to the decoder. This was done by connecting the encoder and the decoder networks through skip connections that created a U-shaped architecture. This U-shaped architecture becomes later a major innovation of FCN and was named “U-Net”. After that, Zhang et al. [65] presented a deep Residual U-Net (ResUNet) that incorporates the strength of both U-Net and residual neural network. Compared to the original U-Net, ResUNet utilized a better CNN backbone that extracted information at multiple scales which caused better performance.

For medical image segmentation, Chen et al. [14] introduced Dense-Res-Inception Net (DRINET) with a better performance when it is compared to FCN, U-Net, and ResUNet. However, employing a dense-inception block increases the growth rate, which may result in too many parameters, making the model more complex and duller to train. Zhou et al. [67] proposed UNet++ for the task of semantic and instance segmentation. The performance of their proposed network was enhanced by restructuring skip connections and developing a pruning strategy for their architecture. They solved the issue of losing edge information and small objects by down-sampling functions. They tested their model on a variety of medical image segmentation tasks.

Additionally, Jha et al. presented (ResUNet++) [34], which is an advanced form of the basic ResUNet. They employed residual blocks, besides integrating additional layers to their network, including squeeze-and-excitation blocks [31], attention blocks, and ASPP [16]. Compared to ResUNet and U-Net, ResUNet++ achieved higher scores in DSC, IoU, and recall. Reza Azad et al. [7] proposed another modification to the conventional U-Net, denoted as Bi-directional ConvLSTM U-Net (BCDU-Net). Besides the full advantages of U-Net, the performance was improved by capturing more discriminative data utilizing bi-directional ConvLSTM and dense convolutions.

Because of the main problem of skip connection, i.e., the issue of the great semantic gap between high- and low-resolution features which results in fuzzy feature maps, Ibtehaz et al. [32] introduced a model to enhance skip connection, titled as MultiResUNet. Their architecture modified the traditional U-Net with Residual Path (ResPath) wherein encoded features execute extra convolution operations before combining the process with equivalent features in the decoder. Asadi-Aghbolaghi et al. [6] presented another U-Net extension for medical image segmentation, named Multi-level Context Gating U-Net (MCGU-Net). In their architecture, they inserted a squeeze-and-excitation (SE) module in the decoder, besides employing BConvLSTM. They utilized a dense convolutions mechanism for extracting richer discriminative features, which led to more fine segmentation maps. Jha et al. [35] proposed the famous DoubleU-Net, which is a blend of two U-Net networks placed on top of one another. On all used datasets, DoubleU-Net has outperformed different baselines and the traditional U-Net. The evaluation was performed on a variety of medical image segmentation tasks.

Zunair et al. [69] introduced Sharp U-Net as an effective depthwise encoder-decoder fully convolutional network for biomedical image segmentation. In their model, they exclude employing skip connection and instead, they utilized sharpening kernel filter in the encoder path. A depthwise convolution of the encoder feature map with a sharpening kernel filter was performed before merging the encoder and decoder features which produced a sharpened intermediate feature map of the same size as the encoder map. A variety

of experiments were performed on six datasets with efficient performance. Azad et al. [9] introduced a new extension to the traditional U-Net, titled a frequency re-calibration U-Net (FRCU-Net). In the skip connection, they employed multi-level BConvLSTM. In addition, they employed SE modules in the decoding path and used densely connected convolutional modules. Moreover, they introduced a frequency-level attention mechanism that utilized a weighted combination of multiple kinds of frequency information to manage and assemble the representation space. However, both FRCU-Net and MCGU-Net need much longer time for convergence because they have a greater number of training parameters. Tran et al. [56] introduced a structural network, titled TMD-Unet. Their model had three major contributions compared to the traditional U-Net. Firstly, they employed three sub-Unet models. Secondly, they utilized dilated convolution (DC) rather than normal convolution. Thirdly, rather than a standard skip connection, they adopted a dense skip connection. A variety of experiments have been performed in this work on different datasets.

3 The proposed Ψ net architecture

The main objective of a supervised network is to learn how to predict the targeted output y from a given input image x , i.e., mapping the input image into the labeled target ($P:x \rightarrow y$), where P is the employed network. The network can learn to extract texture and contextual similarity between the same labeled pixels and the difference between differently labeled neighboring pixels, thereby realistic segmentation can be produced.

Deep learning-based models can provide quick diagnosis and accordingly can support specialists throughout their treatments. Medical image segmentation tasks have mostly adopted U-Net [49] and its related segmentation models [6, 7, 9, 32, 34, 35, 56, 68] to gather both high- and low-level details. U-Net [49], shown in Fig. 3, is a contracting-expanding (encoder-decoder) model which is originally made of a stack of transformers acting as encoder and decoder linked via skip connections. The conventional U-Net is made up of the same number of down-, up-sampling, and convolution layers.

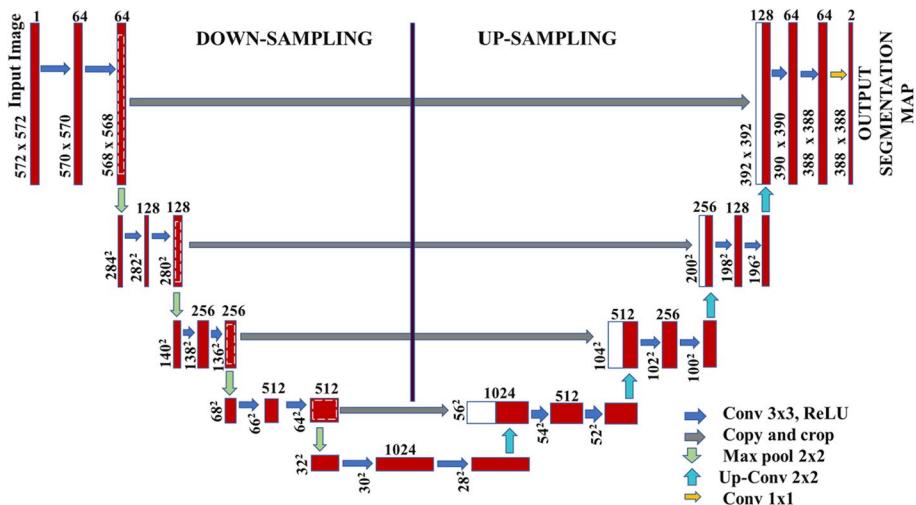


Fig. 3 The schematic architecture of U-Net

Additionally, U-Net connects each pair of down- and up-sampling layers using skip connection operation allowing the spatial information to be straightly transferred to much deeper layers. Hence, highly precise segmentation results can be generated.

Here, we introduce Ψ net as a new deep learning-based segmentation framework. The main goal of the proposed model is to use fewer parameters and, at the same time, maintain high accuracy on a variety of medical image segmentation tasks. The overall view of the proposed Ψ net system is indicated in Fig. 1. The proposed architecture is based on an end-to-end deep learning approach comprised of three U-Net structures. These three U-Nets are connected parallel to each other in which three single U-Nets are fed with input image simultaneously, see Fig. 4. Employing multiple U-Nets helps in capturing more contextual and semantic features efficiently. The proposed segmentation model depends mainly on three parts, i.e., encoder-decoder backbone with squeeze-and-excitation (SE) block, atrous spatial pyramid pooling (ASPP), as well as output module, in a parallel scheme. The proposed segmentation framework has fewer trainable parameters (~33 M) in comparison to FRCU-Net [9] (~68 M), and FCN-8 s [2] (~134 M), which makes it more suitable for real-time performance. In the following subsections, details about the basic building blocks of the proposed segmentation model are demonstrated.

3.1 Encoders

Every encoder tries to encode the input data into representative features at various levels. The encoder increases the channels while decreasing the spatial dimensions in each

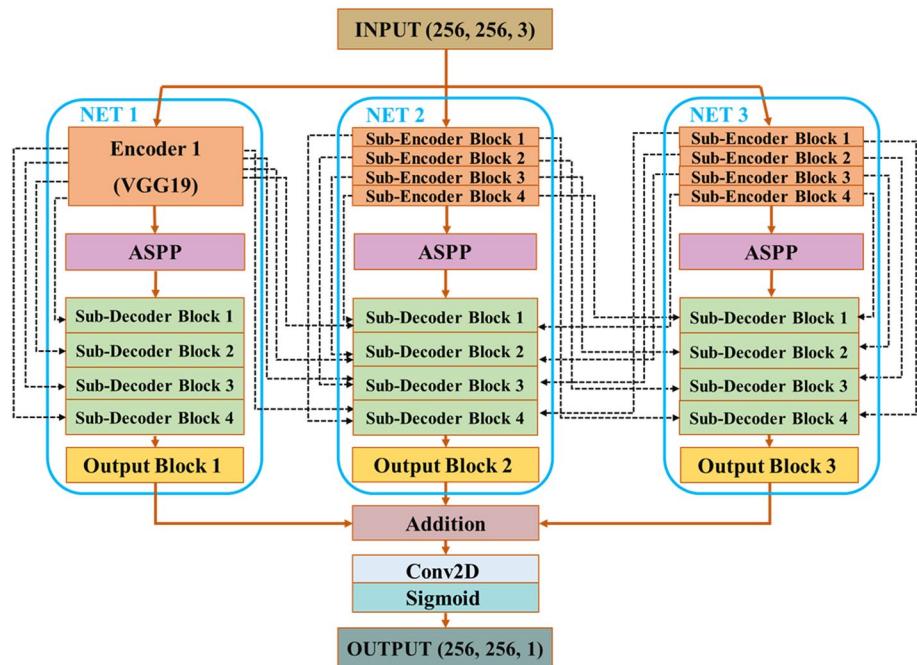


Fig. 4 Detailed schematic structure of the proposed Ψ net. The dashed lines denote the skip connections appended from encoders to decoders

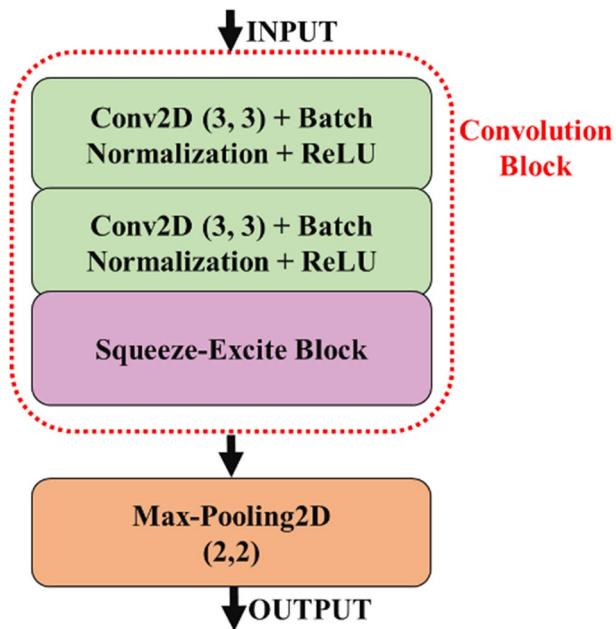
layer. Every encoder in the model receives an input image and its ground truth mask as inputs. A pretrained VGG-19 [51], that has previously been learned on ImageNet features [20], is used in the first encoder of NET 1, while the second and third encoder are built from scratch in NET 2 and NET 3, respectively. The substantial reasons and advantages of adopting VGG-19 are as follows. (1) In comparison to other pre-trained models, it is a lightweight model. (2) VGG-19 and U-Net have similar architecture, which simplifies the integration between them. (3) It provides a suitable deeper network that guarantees a more accurate segmentation mask. VGG-19 has acquired robust feature representations for a diverse set of images.

For encoder 2 and encoder 3 in NET 2 and NET 3, respectively, each one contains four sub-encoder blocks connected serially. Every sub-encoder block includes two main sub-blocks: convolution block and max pooling. The convolution block performs two 3×3 convolution operations, batch normalization, Rectified Linear Unit as an activation function (ReLU), and finally squeeze-excite process, see Fig. 5. For the group of sub-encoders, we employed filter sizes of $\{32, 64, 128, 256\}$. Batch normalization speeds up convergence, decreases internal covariant shift, and regularizes the model. The model non-linearity is represented by a Rectified Linear Unit (ReLU) activation function. The employed squeeze-and-excitation (SE) module promotes feature map quality by increasing their sensitivity towards the main significant features. Finally, to minimize the spatial dimension of the feature maps in the sub-encoder, a max pooling with a 2×2 window and stride 2 is performed.

3.2 Squeeze-and-excitation block (SE)

SE is a representational unit employed to raise the network sensitivity to relevant features and suppress unnecessary ones. SE consists of a global 2D average pooling, two

Fig. 5 The internal structure of each employed sub-encoder block ij inside the encoders of NET 2, and NET 3, where $i \in \{2, 3\}$ denotes the main encoder number and $j \in \{1, 2, 3, 4\}$ denotes the sub-encoder number

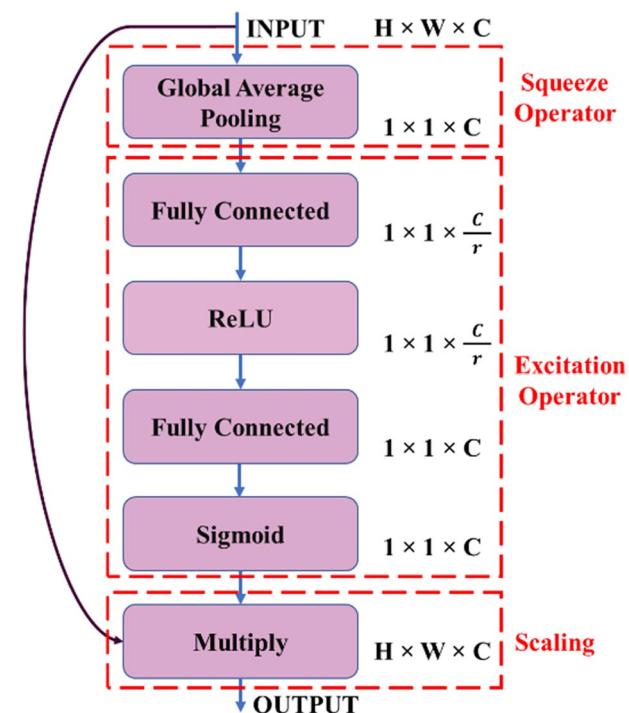


dense blocks, and an element-wise multiplication connected serially. The main target of this block is to weigh every feature map to enhance the representational power of salient features. This objective is accomplished in two phases. The squeezing process is the first phase. It is a global information embedding procedure where every channel is squeezed by utilizing global average pooling to generate channel-wise statistics. Excitation is the second phase in which adaptive recalibration is performed to fully capture channel-wise dependencies. SE block enhances the performance of the network with slight additional computational complexity. It exists inside the convolution block in the proposed network, see Fig. 5. Specifically, we added SE blocks into the intermediate stages inside the convolution block employed in encoder 2, encoder 3, decoder 1, decoder 2, and decoder 3. SE methodology, revise Fig. 6, can be summarized as follows. Firstly, it receives its input from the previous Conv2D layers. Secondly, average pooling is employed to squeeze every channel into a single numerical value. Thirdly, a dense layer followed up by a ReLU provides nonlinearity, while reducing the channel complexity with a ratio r . Then, to provide a smooth gating function to each channel, another dense layer is employed with sigmoid activation. Lastly, every feature map in the convolutional block is weighted according to the “excitation” network. The computational cost of SE modules in the model can be changed by adjusting the reduction ratio (r) as a hyperparameter.

3.3 Atrous spatial pyramidal pooling (ASPP)

ASPP is a resampling module that employs atrous convolutions with different sampling rates to extract multi-scale features [16]. This is performed by applying multiple filters

Fig. 6 A scheme of the employed squeeze-and-excitation (SE) module, where H stands for height, W for width, C for channels, and r denotes the reduction ratio



with different fields of view to the targeted image. Hence, objects and valuable visual context could be captured at different scales. The notion of ASPP is derived from spatial pyramidal pooling [30], which is effective in resampling features at different scales. The ASPP mainly contains two components: atrous convolution and spatial pyramid pooling (SPP). In ASPP, the authors suggested using atrous convolution as an alternative to pooling operation to avoid salient information loss generated by the latter. The atrous convolution can effectively enlarge the field-of-view, i.e., receptive fields of filters, without adding more parameters and can control the resolution of features to extract high-level semantic data. See Fig. 7 for the difference between the regular convolution and the atrous one. Incorporating the advantages of atrous convolution with SPP is proposed by Chen et al. [13] as atrous spatial pyramid pooling (ASPP) module to further boost segmentation performance. ASPP demonstrates high recognition capability on similar objects at multiple levels, which results in a significant accuracy improvement. Hence, ASPP has become a common option in deep segmentation architectures [38], see Fig. 8 for the internal structure of ASPP module. The ASPP block acts as a bridge between the encoder and the decoder in each network branch because it is located in the middle of each branch.

3.4 Decoders

As shown in Fig. 4, we use three decoders in the proposed model corresponding to their three encoders. Each decoder in the proposed model contains four sub-decoder blocks. Each

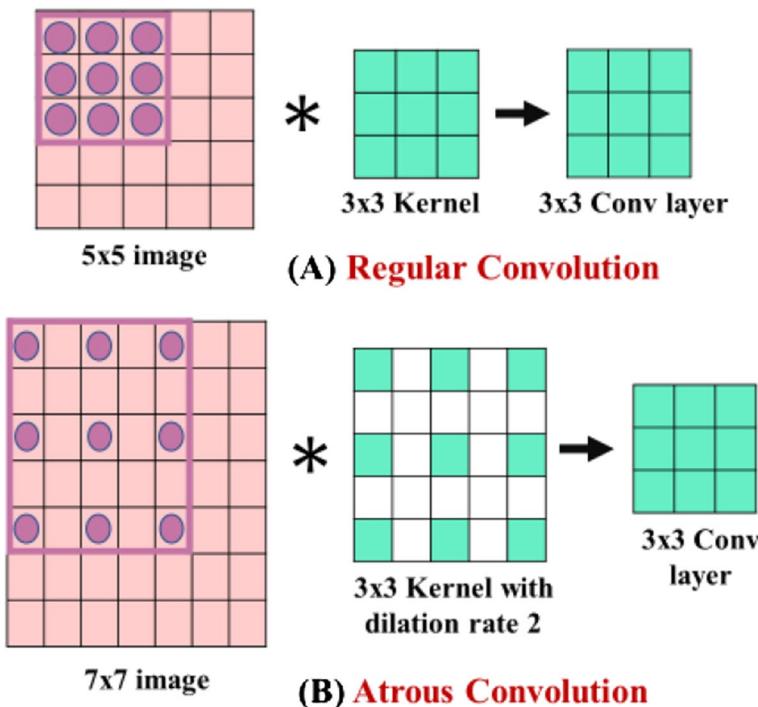


Fig. 7 Illustration of the difference between (a) Regular convolution (Top), and (b) Atrous convolution (Bottom). The number of holes/zeroes filled in between the filter parameters is called the dilation rate

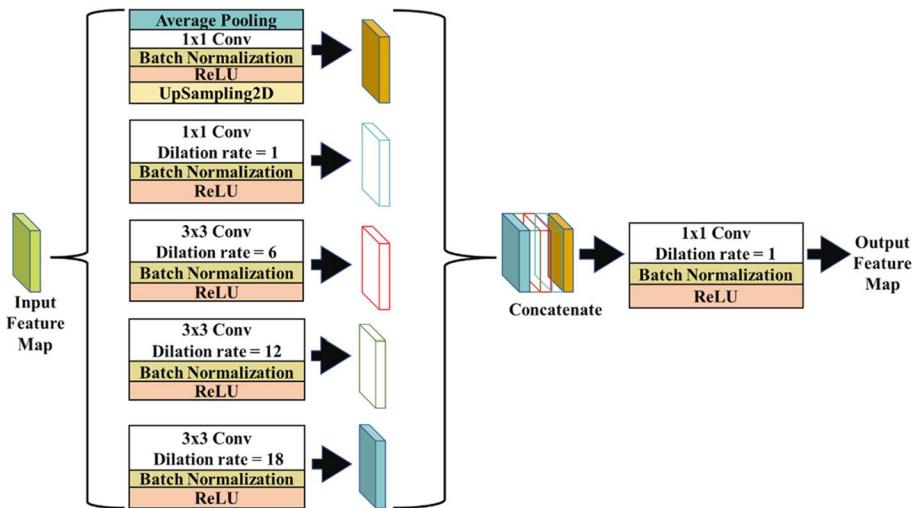


Fig. 8 The internal design of the atrous spatial pyramid pooling (ASPP) module

sub-decoder block doubles the input feature maps dimension by performing a 2×2 bi-linear up-sampling. Then, the group of sub-decoders in NET 1 concatenate the feature maps from the encoder of the same branch through the skip connections and the ASPP block, while the group of sub-decoders in NET 2 and NET 3 concatenate the feature maps from the encoder of the same and the previous branch through the skip connections and the ASPP block. Later, the concatenated maps are applied to the convolution block with filter sizes {256, 128, 64, 32}. Figure 9 represents the internal structure of only one sub-decoder block. All sub-decoders have the same internal layers. The decoders have a special feeding mechanism through skip connections, revise the dashed line in Fig. 4. Skip connections facilitate the gradient flow which leads to easier training and enhances the overall performance of the network. In addition, they help to recover the spatial data wasted due to pooling processes by extracting richer features.

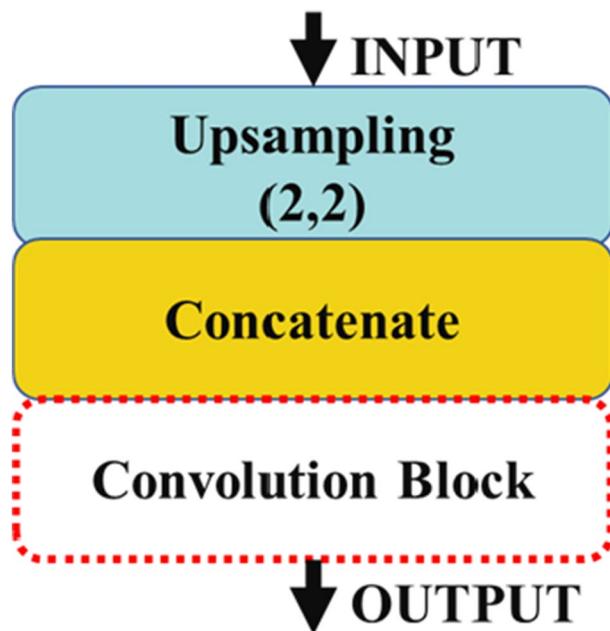
3.5 Output blocks

Finally, as shown in Fig. 4, we use three identical output blocks connected after each decoder block. Each output block applies a convolution layer followed by a sigmoid function. Then, an element-wise addition is performed to the resultant three feature maps. Finally, a 1×1 convolution layer accompanied by a sigmoid activation function is applied to obtain the final segmentation mask.

4 Experimental results and discussion

Experimentally, the proposed segmentation network, i.e., Ψ_{net} , is tested on seven datasets in total, which are Kvasir-SEG [36], CVC-ClinicDB [11], CVC-ColonDB [54], and ETIS-LaribDB [50] for polyp segmentation, 2018 Data Science Bowl challenge [12] for

Fig. 9 The internal structure of the sub-decoder block ij , where $ie \{1, 2, 3\}$ denotes the main decoder number and $je \{1, 2, 3, 4\}$ denotes the sub-decoder number, inside the decoders of NET 1, NET 2, and NET 3



cell nuclei segmentation, ISIC-2017 and ISIC-2018 [18, 19, 57] for skin lesion segmentation. For details about the employed datasets, check Table 1. In addition, Fig. 10 indicates visual samples from the employed datasets.

All experiments were implemented using the Keras framework [17] and TensorFlow [1] as the backend. GPU with high RAM from Google Colab Pro is employed. A batch size of 16 was adopted. We have used dice loss, which is a widely used loss function, based on dice coefficient, for image segmentation tasks [52]. The adaptive moment estimation (Adam) optimizer was utilized with an initial learning rate of $1e-4$ in all experiments. A ReduceLRonPlateau callback, provided by Keras, was employed to monitor the performance. When the validation loss didn't improve after 10 iterations, the learning rate was reduced by a factor of 0.1. In addition, an early stopping was applied when we didn't have any improvements for 20 consecutive epochs. In all datasets, we utilized 80% of the dataset for training, 10% for validation, and the remaining 10% for testing. We trained the model for 300 epochs in all experiments. These hyperparameters were chosen experimentally based on the empirical evaluation.

Evaluation metrics play a pivotal important role in assessing the efficiency of segmentation models, see Table 2 for details about the employed evaluation metrics. In this work, we have analyzed the results using dice similarity coefficient (DSC), Jaccard similarity coefficient (JSC), accuracy (ACC), sensitivity (SE), precision (PREC), and specificity (SPEC) metrics. The empirical results of the proposed methodology on each employed dataset are indicated visually and computationally compared to SOTA techniques. In the performed comparisons, the following segmentation techniques are employed: U-Net [49], ResUNet [65], Recurrent Residual U-Net [4], ResUNet++ [34], UNet++ [68], BCDU-Net [7], MultiResUNet [32], MCGU-Net [6], DoubleU-Net [35], FRCU-Net [9], TMD-Unet [56], Attention Deeplabv3p [8], DDANet [55], and ColonSegNet [37].

Table 1 Details about the employed datasets and their characteristics

Dataset	Task type	Description [Dataset content]	Resolution (w × h)	Organ	Ground Truth	References [Availability]
Kvasir-SEG	Detection, localization, segmentation	1000 polyp images with GT.	Various/Variable resolutions	Large bowel	Polyp regions-binary mask	Debesh Jha et al., 2020 [36]
ETIS-Larib	Segmentation	34 sequences with 44 different polyps yielded 196 White light images containing polyps.	1225×966	Colonoscopy	Polyp locations-binary mask	Silva et al., 2014 [50]
CVC-ClinicDB	Segmentation	612 sequential White light images with polyps obtained from 31 sequences (23 patients) containing 31 different polyps.	384×288	Colonoscopy	Polyp locations-binary mask	Bernal et al., 2015 [11]
CVC-ColonDB	Detection, localization, segmentation	380 sequential White light images with polyps obtained from 15 sequences.	574×500	Colonoscopy	Polyp regions-binary mask	Bernal et al., 2012 [10]; Taibakhsh et al., 2015 [54]
ISIC-2017 Challenge	Segmentation, detection, localization	2000 JPEG lesion images and 2000 associated super-pixel masks.	576×767	Dermoscopy	Skin lesion locations (binary mask)	Codella et al., 2018 [18]
Lesion Boundary Segmentation challenge (ISIC-2018)	Segmentation	12,970 ground truth response masks are related to 2394 images (5 for each image).	700×900	Dermoscopy	contains skin lesion's locations and their corresponding annotations.	Codella et al., 2019 [19]; Tschandl et al., 2018 [57]
2018 Data Science Bowl	segmentation	670 nuclei imaging.	256×256	Microscopy	Nuclei locations (binary mask)	Caicedo et al., 2019 [12]

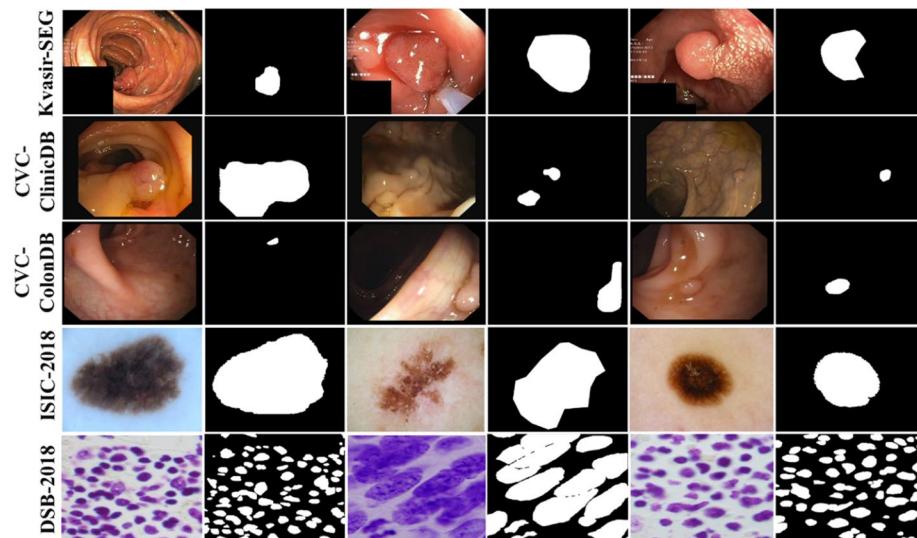


Fig. 10 Samples from the employed datasets with their corresponding ground truth segmentation masks. As indicated, they show variations in shape, size, color, irregular boundaries, and appearance

4.1 Testing the proposed Ψ net on each single/individual dataset

In this subsection, we will indicate the segmentation results obtained after training and testing the suggested segmentation network on a single dataset each time.

4.1.1 Skin lesions segmentation

For the task of segmenting skin lesions, two well-known dermoscopic imaging datasets are used, i.e., ISIC (International Skin Imaging Collaboration) challenge 2017, and 2018. The ISIC database comprises RBG dermoscopy images and their corresponding ground truth binary masks to designate lesion boundaries. Due to the lesions' diverse characteristics in terms of shapes, colors, and textures, this segmentation task is extremely challenging.

A. ISIC-2017 Challenge The ISIC 2017 competition [18] consists of three challenges: lesion segmentation, dermoscopic feature detection, and disease classification. Here, we target the task of segmentation. The proposed Ψ net is compared to the traditional U-Net [49] besides the most recent studies in skin lesion segmentation. Table 3 depicts the quantitative results of the performed comparison. As indicated, the proposed Ψ net shows the top F1-score performance with an approximate improvement margin of 1.25% compared to MCGU-Net [6] which shares the same idea of inserting the SE module in the decoder. In addition, the introduced method outperforms the commonly used BCDU-Net [7], which employs a bidirectional scheme in their enhancement strategy, in terms of F1-score with a 2.42% increase. On the other side, a 3.7% improvement in F1-score is achieved compared to the traditional U-Net [49]. In terms of specificity, the Ψ net model outperforms MCGU-Net [6]. On the other side, the proposed Ψ net comes in second place after MCGU-Net [6], in terms of accuracy, with a slight difference but it still outperforms the traditional U-Net

Table 2 The employed evaluation metrics, where GT indicates Ground Truth, SR represents Segmentation Result, TP denotes True Positive, TN denotes True Negative, FP for False Positive, FN for False Negative

Metric	Equation	Definition
Accuracy (ACC)	$\frac{TP+TN}{TP+TN+FP+FN}$	- Accuracy measures the number of correct predictions attributed to all predictions made.
Specificity (SPEC)	$\frac{TN}{\# \text{ of all negatives}} = \frac{TN}{TN+FP}$	- Specificity calculates the ratio of true negative rates (TNR).
Precision (PREC)	$\frac{TP}{TP+FP}$	<ul style="list-style-type: none"> - Precision accurately describes the purity of positive detections in comparison to ground truth. This metric is frequently called the positive predictive value. - Precision is a good metric to be considered when the costs of false positives are substantial. - Recall determines the true positive rate (TPR). It accurately depicts the completeness of our positive expectations in comparison to ground truth. - Recall, as well, called sensitivity (SE), is the proportion of correctly detected positive instances to all actual positive ones, which are the total of “True Positives” and “False Negatives”.
Recall (REC)	$\frac{TP}{TP+FN}$	<ul style="list-style-type: none"> - F1-score is a statistical metric known as the harmonic mean of precision and recall.
F1-score	$\frac{2 * Precision * Recall}{Precision + Recall} = \frac{2TP}{2TP+FP+FN}$	- Dice score (DSC) is related to precision in that it penalizes for false positives that the method uncovers. In addition, it measures how many positives are discovered. Hence, it is closer to precision than accuracy.
DSC	$\frac{2 GT \cap SR }{ GT + SR } = \frac{2TP}{2TP+FP+FN}$	<ul style="list-style-type: none"> - DSC calculates the overlap of whole regions between ground-truth and predicted labels. It is ranged from 0 to 1, where 1 denotes a perfect and total overlap. It is equivalent to F1-score.
IoU / JSC	$DSC = \frac{2 * IoU}{IoU+1}$ $IoU = \frac{ GT \cap SR }{ GT \cup SR } = \frac{TP}{TP+FP+FN}$ $IoU = \frac{DSC}{2-DSC}$	<ul style="list-style-type: none"> - The Jaccard similarity index (JSC) / Intersection over Union (IoU) calculates the percentage of overlap between the prediction output and target mask. This metric is highly tied to DSC, which is frequently employed during training as a loss function. - In other words, IoU calculates the pixels shared by the target and prediction masks, divided by the total pixels included in both masks. It focuses more on overlapping regions of two sets.

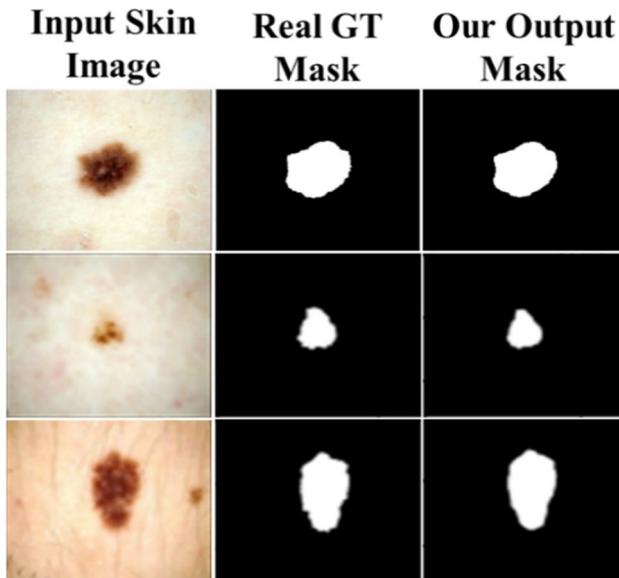
- TP (True Positive): pixels that were properly segmented as foreground
- TN (True Negative): pixels that were correctly detected as background
- FP (False Positive): pixels that were incorrectly segmented as foreground
- FN (False Negative): pixels that were falsely detected as background

Table 3 Computitative comparison between the proposed model and the most common segmentation models on ISIC-2017. The best results are indicated in bold font, while the second place is underlined

Methods	F1-score	REC	SPEC	ACC	IoU
U-Net [49]	0.8682	0.9479	0.9263	0.9314	0.9314
Codella, Melanoma det. [18]	–	–	–	0.9340	–
Al-Masni, Lesion Analysis [3]	–	0.8250	0.9750	0.9340	–
Zahangir Alom, R2U-net [63]	0.8920	<u>0.9414</u>	0.9425	0.9424	0.9421
BCDU-Net [7]	0.8810	0.8647	0.9751	0.9528	<u>0.9528</u>
MCGU-Net [6]	<u>0.8927</u>	0.8502	<u>0.9855</u>	0.9570	0.9570
Ψnet (ours)	0.9052	0.8537	0.9887	<u>0.9538</u>	0.8268

[49] by an approximate accuracy margin of 2.24%. However, the traditional U-Net [49] is still the top performer in terms of recall, while MCGU-Net [6] is the top performer in terms of IoU. For supporting visual results, see Fig. 11. In addition, Fig. 12 demonstrates output segmentation masks using the proposed methodology compared to other SOTA methods.

B. ISIC-2018 Challenge This dataset [19] is a large-scale dataset of dermoscopy images that the International Skin Imaging Collaboration (ISIC) published in 2018. For a quantitative comparison between the proposed network and some common different alternatives, see Table 4. From this quantitative comparison, the suggested network achieves better performance compared to SOTA alternatives in terms of F1-score, sensitivity, accuracy, and precision. The proposed Ψ net shows superior performance compared to FRCU-Net [9], which includes a multi-level BConvLSTM and SE, by achieving an F1-score improvement of 1.8%. In addition, the proposed network overcomes the commonly known

**Fig. 11** Visual segmentation results of the proposed model on ISIC-2017

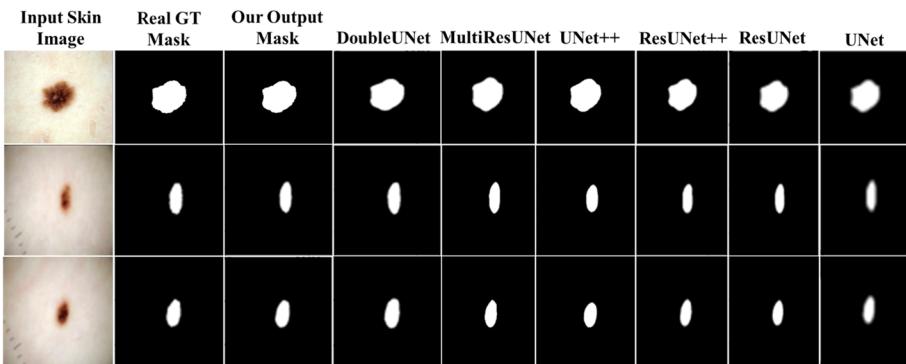


Fig. 12 Visual outputs using the proposed methodology compared to various SOTA models on ISIC-2017

BCDU-Net [7], which utilizes BConvLSTM and dense convolutions, with an 8% F1-score improvement. Moreover, superior and huge improvements, i.e., 4.7% and 28.4%, as well as 9.4% and 32.1%, are achieved compared to TMD-Unet [56], traditional U-Net [49] in terms of F1-score and IoU, respectively. Furthermore, the proposed Ψ net with Attention Deeplabv3+ [8] shows the highest accuracy of 0.964 compared to the other alternatives with an approximate margin of 7.4% improvement compared to U-Net [49]. However, in terms of specificity, the proposed Ψ net model, TMD-Unet [56], and BCDU-Net [7] achieve the third top score of 0.982 after Attention Deeplabv3+ [8] and MCGU-Net [6] with a very slight difference. For some qualitative segmentation masks of the proposed network on the ISIC-2018 dataset, see Fig. 13. As indicated, the proposed network performs efficiently on all kinds of lesions from small to large ones. In addition, Fig. 14 shows different output segmentation results using the proposed methodology compared to other SOTA methods.

Table 4 Quantitative comparison on skin lesion segmentation challenge ISIC-2018 between the proposed Ψ net and the most common segmentation models. The best results are bolded, and the second place is underlined

Methods	F1-score	REC	SPEC	ACC	IoU	PREC
U-Net [49]	0.647	0.708	0.964	0.890	0.549	0.779
Attention U-Net [44]	0.665	0.717	0.967	0.897	0.566	0.787
R2U-Net [63]	0.679	0.792	0.928	0.880	0.581	0.741
Attention R2U-Net [63]	0.691	0.726	0.971	0.904	0.592	0.822
BCDU-Net [7]	0.851	0.785	0.982	0.937	0.937	0.928
MCGU-Net [6]	0.895	0.848	<u>0.986</u>	0.955	0.955	<u>0.947</u>
Deeplabv3+ [15]	0.882	0.856	0.977	0.951	0.951	–
Attention Deeplabv3+ [8]	0.912	0.875	0.988	0.964	0.964	–
FRCU-Net [9]	<u>0.913</u>	<u>0.904</u>	0.979	<u>0.963</u>	<u>0.963</u>	0.922
TMD-Unet [56]	0.884	0.850	0.982	–	0.776	0.925
Ψnet (ours)	0.931	0.913	0.982	0.964	0.870	0.948

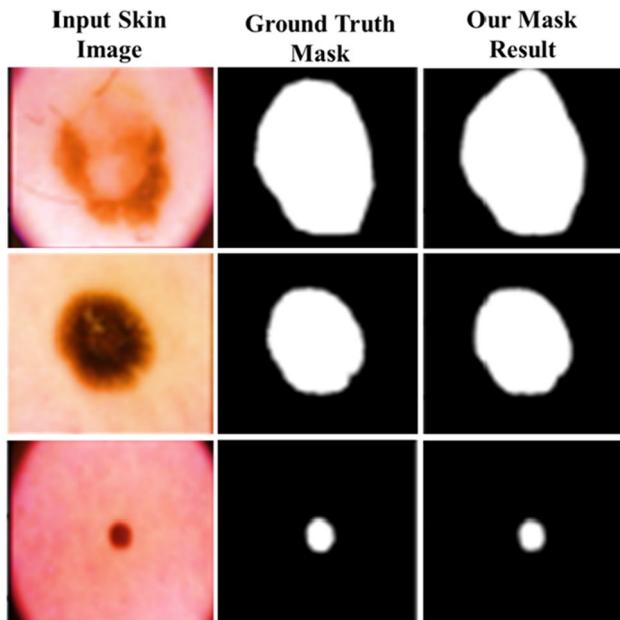


Fig. 13 Some visual segmentation results using the proposed Ψ net on ISIC-2018

4.1.2 Polyps segmentation

Colonoscopy is an efficient mechanism to expose colorectal polyps that are closely associated with colorectal cancer. Segmenting polyps from colonoscopy images is crucial in clinical practice because it gives important information for diagnosis and surgery. However, the appearances, sizes, colors, textures, and aspect ratios of polyps in colonoscopy images vary, even of the same type. In addition, there is no sharp boundary between a polyp and the mucosa around it. Hence, precise polyp segmentation is a difficult task.

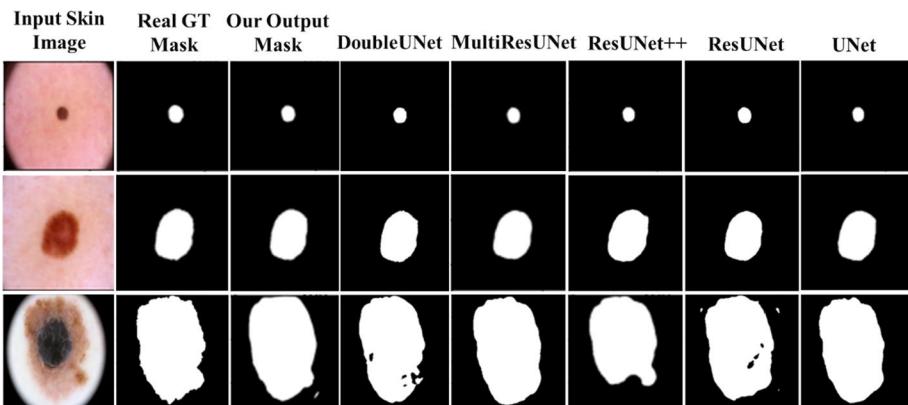


Fig. 14 Visual results of the proposed model compared to other SOTA models on ISIC-2018 dataset

Table 5 Quantitative comparison of our model and the SOTA polyp segmentation methods on Kvasir, ClinicDB, ColonDB, and ETIS

Methods	Kvasir		ClinicDB		ColonDB		ETIS	
	DSC	IoU	DSC	IoU	DSC	IoU	DSC	IoU
U-Net [49]	0.818	0.746	0.823	0.755	0.512	0.444	0.398	0.335
UNet++ [68]	0.821	0.743	0.794	0.729	0.483	0.410	0.401	0.344
SFA [25]	0.723	0.611	0.700	0.607	0.469	0.347	0.297	0.217
PraNet [24]	0.898	0.840	0.899	0.849	0.709	0.640	0.628	0.567
MSNET [66]	0.907	0.862	0.921	0.879	0.755	0.678	0.719	0.664
Shallow Attention [60]	0.904	0.847	0.916	0.859	0.753	0.670	0.750	0.654
PlutoNet [22]	0.894	0.808	0.923	0.857	0.718	0.560	0.748	0.597
Ψnet (ours)	0.904	0.827	0.944	0.895	0.926	0.864	0.888	0.800

For the evaluation of Ψ net, we conducted experiments on four common colonoscopy polyp segmentation benchmarks, i.e., Kvasir-SEG, CVC-ColonDB, CVC-ClinicDB, and ETIS-Larib. The achieved performance is better than most of SOTA methods. Despite the prementioned challenges in polyp segmentation, the proposed Ψ net shows effectiveness in segmentation, check Table 5 for the results of all utilized colonoscopy datasets. The next subsections indicate more detailed quantitative and qualitative results of each polyp dataset which prove the accuracy and the generalizability of the proposed Ψ net.

A. Kvasir-SEG Challenge The first used colonoscopy dataset in the performed experiments is Kvasir-SEG dataset [36]. It is publicly available for polyp detection, localization, and segmentation. Sample images from Kvasir-SEG dataset and their corresponding masks are displayed in Fig. 10. The quantitative evaluation of Ψ net is reported in Table 6, while the supporting qualitative results are shown in Figs. 15 and 16. As indicated, the proposed model shows superior performance in all metrics compared to ResUNet, ResUNet++, NanoNet, and DDA Net. Compared to DDA Net, the proposed methodology shows superior performance by achieving improvements of 4.69%, 4.7%, and 7.68% in terms of DSC, IoU, and precision, respectively. In addition, Ψ net achieved an approximate increase of 8.09% in terms of recall compared to ResUNet++, which is based on encoder-decoder architecture with residual and SE blocks. The proposed framework ability to segment polyps can

Table 6 Quantitative comparison of the Kvasir-SEG dataset between the proposed Ψ net and the most common SOTA polyp segmentation models. The best scores are bolded, while the second place is underlined

Methods	DSC	IoU	REC	PREC	ACC
ResUNet (GRSL'18) [65]	0.7203	0.6106	0.7602	0.7624	0.9251
ResUNet++ (ISM'19) [34]	0.7310	0.6363	0.7925	0.7932	0.9223
NanoNet-A [38]	0.8227	0.7282	0.8588	0.8367	0.9456
NanoNet-B [38]	0.7860	0.6799	0.8392	0.8004	0.9365
NanoNet-C [38]	0.7494	0.6360	0.8081	0.7738	0.9290
DDANet [55]	<u>0.8576</u>	<u>0.7800</u>	0.8880	<u>0.8643</u>	–
Ψnet (SE with $r=8$)	0.9045	0.8270	<u>0.8734</u>	0.9411	0.9633
Ψnet (SE with $r=16$)	0.9012	0.8216	0.8770	0.9369	0.9629

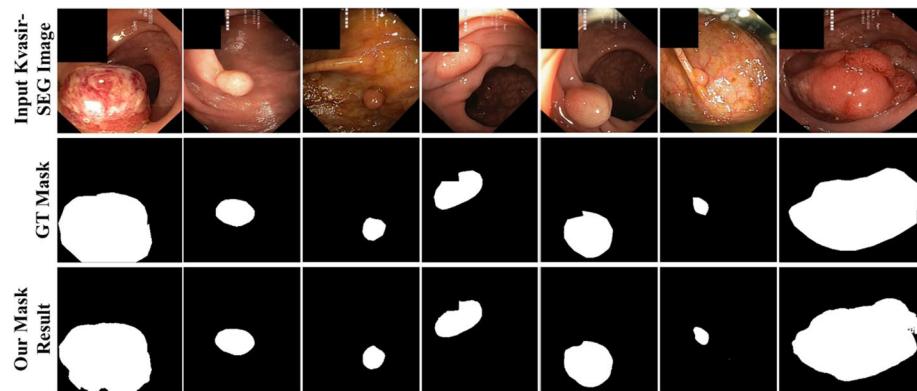


Fig. 15 Visual qualitative segmentation results of Ψ net on medium, flat, and large polyps from the Kvasir-SEG dataset for automatic polyp detection

be seen by comparing the ground truth to the predicted mask, see Fig. 15. Furthermore, Fig. 16 depicts the segmentation outputs using the proposed methodology compared to other SOTA techniques. From the visual results, we can deduce that the proposed model generates better remarkable results with small- and large-sized polyps.

B. CVC-ClinicDB The second employed colonoscopy dataset is CVC-ClinicDB [11], known as CVC-612. Based on the quantitative results in Table 7, the proposed model achieves the highest DSC, IoU, and precision compared to other SOTA methods, such as U-Net, Deeplabv3+(Xception), Deeplabv3+(Mobilenet), HRNetV2-W18-Smallv2, HRNetV2-W48, ResUNet++, ResUNet++ + CRF, DoubleU-Net, and TMD-Unet. The DSC and IoU scores are important metrics in segmentation tasks. ResUNet++ + CRF and ResUNet++ achieve the highest recall scores with a very small difference between each other. Compared to DoubleU-Net, there is a large gap in the recall by approximately 6.97% from ours. In addition, the proposed Ψ net surpasses the baseline architectures, like U-Net and Deeplabv3+(Xception), with a significant margin, in terms of IoU and DSC, with

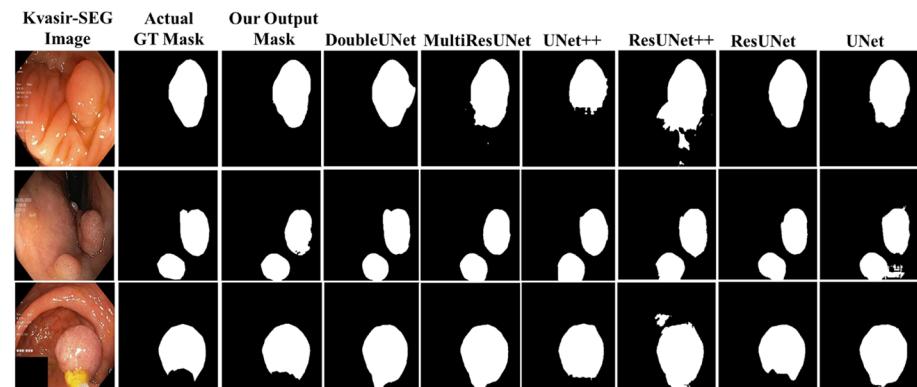


Fig. 16 Visual comparison of the proposed model on the Kvasir-SEG dataset against various SOTA models

Table 7 Quantitative comparison between the proposed Ψ net and the most common SOTA segmentation models on CVC-ClinicDB. The best scores are bolded, and the second place is underlined

Methods	DSC	IoU	REC	PREC
U-Net [49]	0.8781	0.7881	0.7865	0.9329
Deeplabv3+(Xception) [15]	0.8897	0.8706	0.9251	0.9366
Deeplabv3+(Mobilenet) [15]	0.8985	0.8588	0.9160	0.9287
HRNetV2-W18-Smallv2 [59]	0.1337	0.0867	0.1391	0.1854
HRNetV2-W48 [59]	0.1924	0.1284	0.2055	0.3042
ResUNet++ [34]	0.9199	0.8892	<u>0.9391</u>	0.8445
PraNet [24]	0.8990	0.8490	—	—
ResUNet++ + CRF [39]	0.9203	<u>0.8898</u>	0.9393	0.8459
DoubleU-Net [35]	0.9239	0.8611	0.8457	<u>0.9592</u>
TMD-Unet [56]	<u>0.9265</u>	0.8641	0.9225	0.9376
Ψnet (SE with $r=8$)	0.9449	0.8958	0.9154	0.9665
Ψnet (SE with $r=16$)	0.9502	0.9052	0.9270	0.9663

10.77%, 2.52% and 6.68%, 5.52% improvements, respectively. Moreover, it outperforms existing SOTA techniques, like DoubleU-Net, and TMD-Unet, and achieves a higher DSC of 0.9449. Comparing our methodology to ground truth masks, the predicted masks have substantially identical polyp boundaries and shapes as shown in Fig. 17. In addition, some output segmentation masks using the proposed model in comparison to some SOTA methods are indicated in Fig. 18.

C. CVC-ColonDB CVC-ColonDB [54] is the third polyp dataset that is employed for a more comprehensive performance analysis of automatic polyp segmentation. The quantitative results in Table 8 demonstrate that the proposed Ψ net outperforms other SOTA techniques with a superior DSC of 0.9269 by an approximate 7.95% improvement compared to ResUNet++ + TTA. There is a slight difference in DSC values between ResUNet++ and ResUNet++ + CRF. In addition, superior performance is reported in terms of IoU of 0.8641 with an approximate 1.75% improvement compared to ResUNet++ + TTA. In terms of recall, the proposed network demonstrates better performance compared to ResUNet++

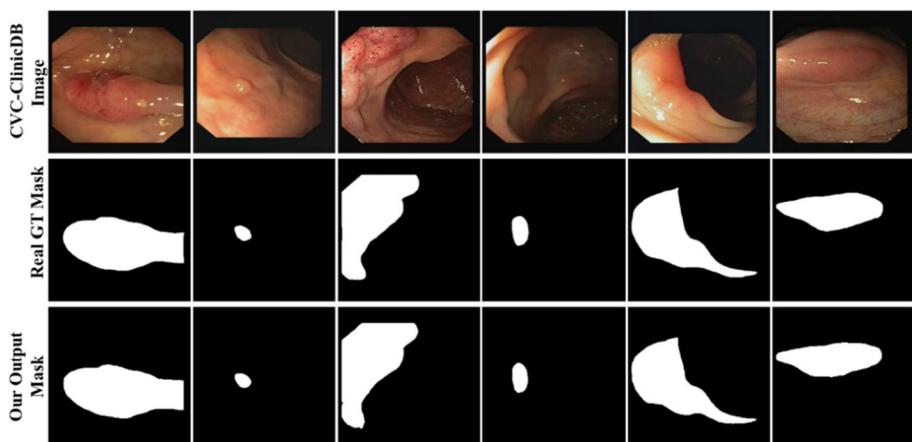


Fig. 17 Visual segmentation results of the proposed Ψ net on CVC-ClinicDB

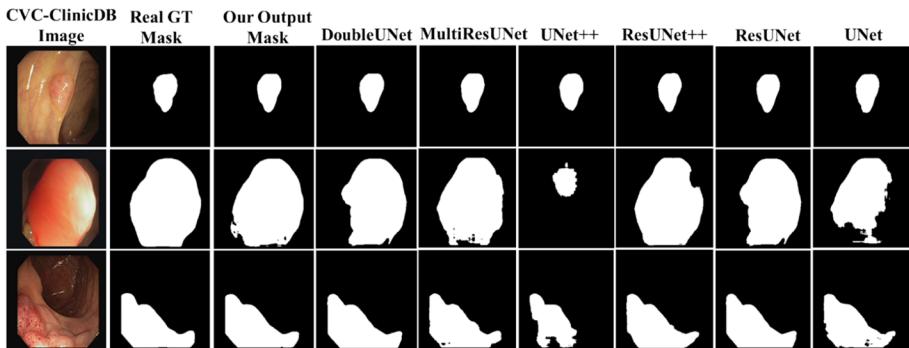


Fig. 18 Visual comparison of the proposed Ψ net and various SOTA models on CVC-ClinicDB

and ResUNet++ + CRF with 9.12%, and 9.26% enhancement, respectively. Moreover, it achieves a higher score in precision by an approximate 11.75% increase compared to ResUNet++ + TTA + CRF. From the visual qualitative outputs in Fig. 19, we can conclude that the developed network provides exact segmentation masks when compared to ground truth masks. More visual results are indicated in Fig. 20 compared to other SOTA methods.

D. ETIS-Larib Dataset ETIS-LaribPolypDB [50] is the fourth employed polyp dataset. It is the most challenging one because the majority of polyps are small and hard to identify. The proposed network achieved the best DSC, IoU, precision, and recall compared to some SOTA methods, such as PraNet, ResUNet++, ResUNet++ + CRF, ResUNet++ + TTA, and ResUNet++ + TTA + CRF, as indicated in Table 9. It achieves the highest DSC and IoU of 0.8888 and 0.8000 with 25.24% and 4.66% improvements, respectively, compared to ResUNet++. Moreover, it provides the highest precision with 0.9767. Furthermore, it achieves 26.4% and 33% improvements in recall and precision, respectively, compared to ResUNet++, and improvements of 29.9% and 32.02% compared to ResUNet++ + TTA.

ETIS-LaribPolypDB is the most dataset affected by the changes in the reduction ratio r inside the SE block. By changing r from 8 to 16, a huge enhancement is achieved in all

Table 8 Quantitative comparison between the proposed Ψ net and the most common SOTA segmentation models on CVC-ColonDB. The best scores are bolded, and the second place is underlined

Methods	DSC	IoU	REC	PREC
FCN-8S + Otsu [2]	0.8100	–	0.7480	–
FCN-8 s + Texon [64]	0.7014	–	0.7566	–
SA-DOVA Descriptor [10]	0.5533	–	0.6191	–
PraNet [24]	0.7090	0.6400	–	–
ResUNet++ [34]	0.8469	0.8456	<u>0.8511</u>	0.8003
ResUNet++ + CRF [39]	0.8458	0.8456	0.8497	0.7767
ResUNet++ + TTA [39]	<u>0.8474</u>	<u>0.8466</u>	0.8434	0.8118
ResUNet++ + TTA + CRF [39]	0.8452	0.8459	0.8411	<u>0.8125</u>
Ψnet (ours), (SE with $r=8$)	0.9269	0.8641	0.9423	0.9300
Ψnet (ours), (SE with $r=16$)	0.9299	0.8693	0.9359	0.9295

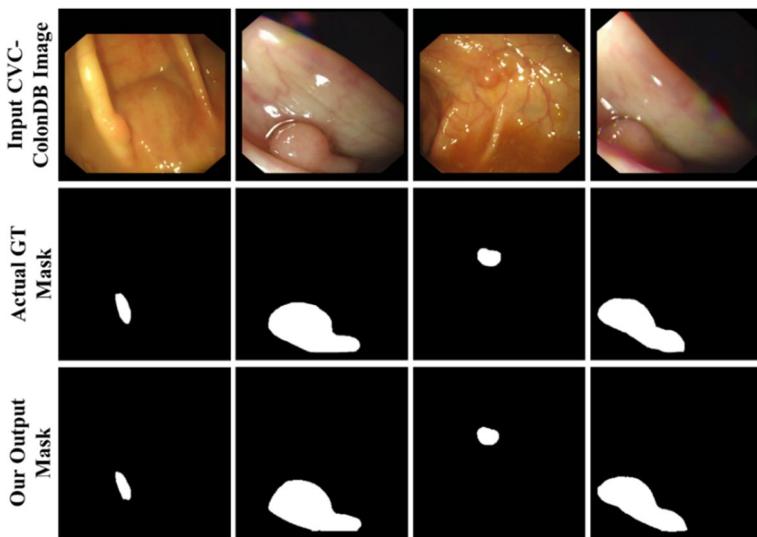


Fig. 19 Visual segmentation results of Ψ net on challenging images from CVC-ColonDB

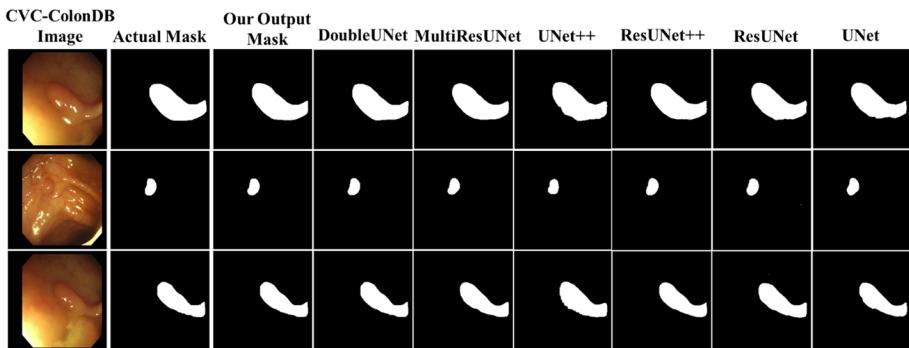


Fig. 20 Visual qualitative comparison on CVC-ColonDB of the proposed model and various SOTA models

Table 9 Quantitative comparison between the proposed Ψ net and the most common SOTA polyp segmentation models on ETIS-LaribPolypDB. The best scores are bolded, and the second place is underlined

Methods	DSC	IoU	REC	PREC
PraNet [24]	0.6280	0.5670	—	—
ResUNet++ [34]	<u>0.6364</u>	<u>0.7534</u>	<u>0.6346</u>	0.6467
ResUNet++ + CRF [39]	0.6228	0.7520	0.6242	0.5648
ResUNet++ + TTA [39]	0.6136	0.7458	0.5996	<u>0.6565</u>
ResUNet++ + TTA + CRF [39]	0.6018	0.7426	0.5914	0.5755
Ψnet (ours), (SE with $r=8$)	0.8888	0.8000	0.8986	0.9767
Ψnet (ours), (SE with $r=16$)	0.9343	0.8768	0.9118	0.9771

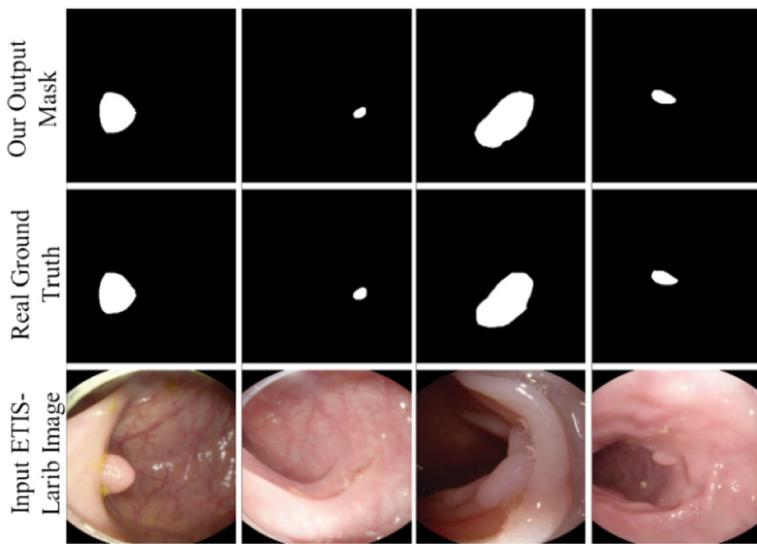


Fig. 21 Some visual segmentation results of Ψ net on ETIS-LaribPolypDB

metrics with a 4.55% increase in DSC, 7.68% in IoU, 1.32% in recall, and 0.04% in precision. In addition, there is a high difference between the performance of the proposed Ψ net and the other listed techniques that makes our model a new strong baseline for medical image segmentation. For supporting visual results, see Figs. 21 and 22. Therefore, the visual and computational results reveal the significance of the proposed network in providing automated polyp detection and delineation with less miss-detection rates.

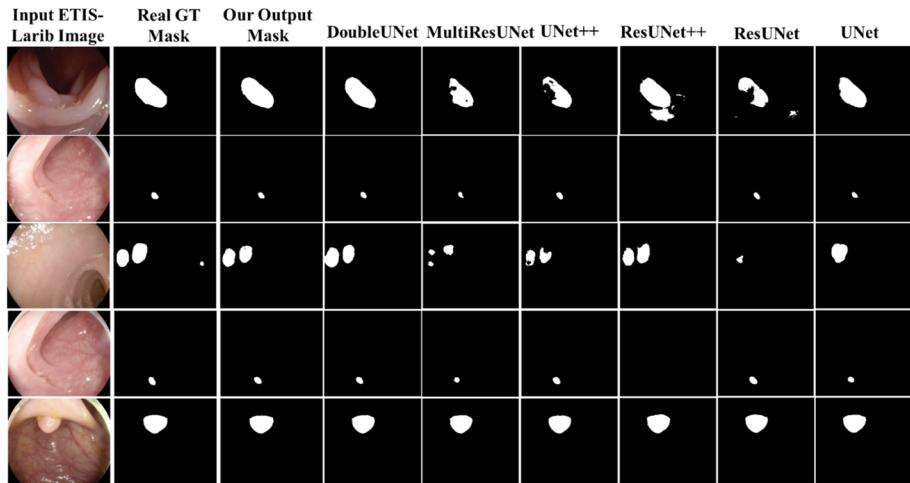


Fig. 22 Visual comparison of the proposed model and other several SOTA models on ETIS-LaribPolypDB

4.1.3 Nuclei segmentation

Nuclei segmentation is a key technique for automatic pathological screening. Precisely segmented nuclei are necessary, not only for cancer detection but for defining the proper treatment effectively. Nevertheless, the variety in cell types and sizes, as well as the variety in extrinsic influences, and illumination circumstances, make nucleus segmentation a difficult task. For nucleus segmentation, the 2018 Data Science Bowl (DSB) dataset [12] is employed which is publicly available at Broad Bioimage Benchmark Collection (<https://data.broadinstitute.org/bbbc/BBBC038/>).

The goal of the DSB 2018 challenge is to find nuclei in divergent images. The quantitative results in Table 10 demonstrate that the proposed Ψ net outperforms other networks, such as U-Net, UNet++, ResUNet++, Deeplabv3+ (Xception), Deeplabv3+ (Mobilenet), HRNetV2-W18-Smallv2, HRNetV2-W48, ResUNet++ + CRF, PraNet, and TMD-Unet, in terms of DSC, IoU, and precision. In addition, the proposed approach achieves the top scores in DSC, IoU, and Precision of 0.9243, 0.8632, and 0.9252 with 0.87%, 1.51%, and 2.83% improvements, respectively, compared to TMD-Unet. ColonSegNet is the top precision performer while TMD-Unet is the top recall performer. The developed model surpasses ResUNet++ in terms of DSC and IoU with 1.45% and 2.62% improvements, respectively. In addition, it achieves an increase of 1.1% enhancement in recall compared to PraNet. For supporting visual results, see Figs. 23 and 24 which indicate some output segmentation masks using the proposed methodology compared to various SOTA methods.

4.2 Cross-testing the employed four colonoscopy imaging datasets

Usually, the generalization capability of a specific model is evaluated by performing a blind test on a part of the employed dataset that the model did not see before. More generalization can be tested by evaluating its applicability across various datasets from multiple sources. Cross-data evaluation is important to validate the model on unseen polyps from other sources. Hence, in this subsection, we will indicate the segmentation results of employing the proposed segmentation network after a training process on a single specific polyp dataset while the testing is performed on other datasets besides the training one.

Table 10 Quantitative comparison between the proposed Ψ net and the most common SOTA segmentation models on 2018 Data Science Bowl dataset. The highest results are bolded, while the second place is underlined

Methods	DSC	IoU	REC	PREC
U-Net [49]	0.9080	0.8314	0.9029	0.9130
UNet++ [67]	0.7705	0.5265	0.7159	0.6657
ResUNet++ [34]	0.9098	0.8370	0.9169	0.9057
Deeplabv3+ (Xception) [15]	0.8857	0.8367	0.9141	0.9081
Deeplabv3+ (Mobilenet) [15]	0.8239	0.7402	0.8896	0.8151
HRNetV2-W18-Smallv2 [59]	0.8495	0.7585	0.8640	0.8398
HRNetV2-W48 [59]	0.8488	0.7588	0.8359	0.8913
ColonSegNet [37]	<u>0.9197</u>	0.8466	0.9153	0.9312
ResUNet++ + CRF [39]	0.7806	0.7322	0.7534	0.6308
PraNet [24]	0.8751	0.7868	0.9182	0.8438
TMD-Unet [56]	0.9156	<u>0.8481</u>	0.9422	0.8969
Ψnet (ours), (SE with $r=8$)	0.9243	0.8632	<u>0.9292</u>	<u>0.9252</u>
Ψnet (ours), (SE with $r=16$)	0.9245	0.8630	0.9221	0.9323

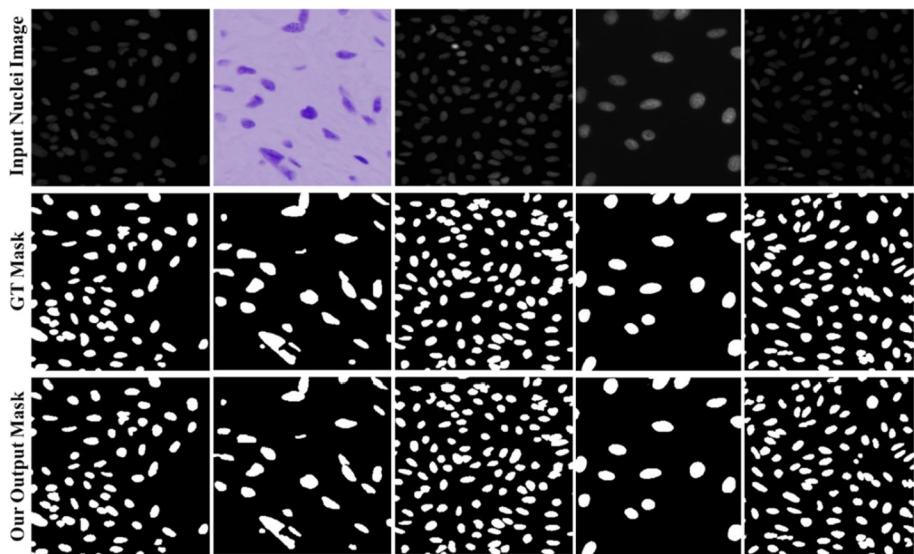


Fig. 23 Visual segmentation results of Ψ net on 2018 Data Science Bowl

Detailed results are indicated in Tables 11, 12, 13 and 14 and demonstrated in the following subsections.

4.2.1 CVC-ClinicDB-based cross-evaluation

Here, the model is trained on CVC-ClinicDB, then the testing process is performed on CVC-ClinicDB, besides the other three polyp datasets, i.e., CVC-ColonDB, ETIS-Larib, and Kvasir-SEG. Table 11 shows the results of cross-testing the proposed Ψ net. As

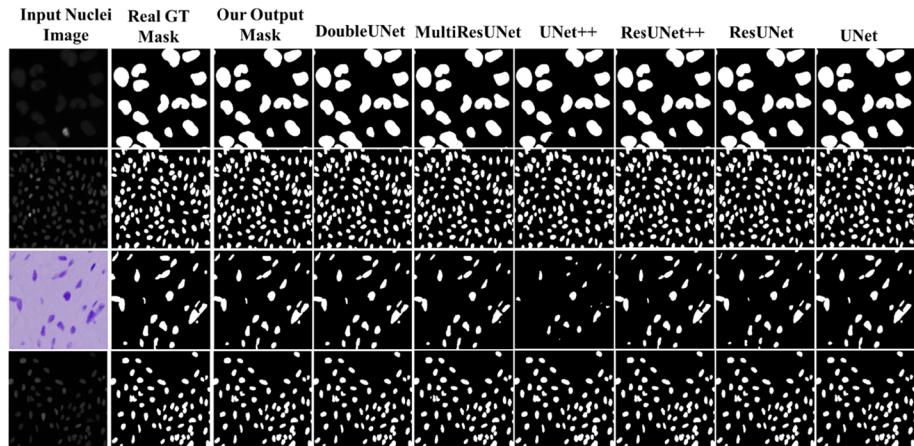


Fig. 24 Visual qualitative results of the proposed model on 2018 Data Science Bowl compared to various SOTA models

Table 11 Cross-testing results, where Ψ_{net} is trained on CVC-ClinicDB and tested on Kvasir-SEG, CVC-ColonDB, ETIS-Larib, and CVC-ClinicDB. The highest scores are bolded, and the second place is underlined

Training dataset	Testing datasets	Metrics					
		DSC	ACC	PREC	REC	IoU	loss
CVC-ClinicDB	CVC-ClinicDB	0.9449	0.9865	0.9665	0.9154	0.8958	0.0551
	CVC-ColonDB	0.6494	0.9585	0.8870	0.5207	0.4953	0.3506
	ETIS-Larib	0.7373	<u>0.9791</u>	0.8848	<u>0.6730</u>	0.5893	<u>0.2627</u>
	Kvasir-SEG	<u>0.7688</u>	0.9300	<u>0.9486</u>	0.6212	<u>0.6299</u>	0.2312

Table 12 Cross-testing results, where Ψ_{net} is trained on Kvasir-SEG and tested on CVC-ClinicDB, CVC-ColonDB, ETIS-Larib, and Kvasir-SEG. The highest scores are bolded, and the second place is underlined

Training dataset	Testing datasets	Metrics					
		DSC	ACC	PREC	REC	IoU	loss
Kvasir-SEG	CVC-ClinicDB	<u>0.8120</u>	<u>0.9675</u>	<u>0.9198</u>	0.7400	<u>0.6889</u>	0.1893
	CVC-ColonDB	0.7215	0.9653	0.8784	0.6400	0.5762	<u>0.2799</u>
	ETIS-Larib	0.6433	0.9687	0.6346	<u>0.7638</u>	0.4941	0.3334
	Kvasir-SEG	0.9045	0.9633	0.9411	0.8734	0.8270	0.0971

Table 13 Cross-testing results, where Ψ_{net} is trained on CVC-ColonDB and tested on CVC-ClinicDB, CVC-ColonDB, ETIS-Larib, and Kvasir-SEG. The highest scores are bolded, and the second place is underlined

Training dataset	Testing datasets	Metrics					
		DSC	ACC	PREC	REC	IoU	loss
CVC-ColonDB	CVC-ClinicDB	<u>0.6075</u>	0.9426	0.9486	<u>0.4277</u>	<u>0.4435</u>	0.3976
	CVC-ColonDB	0.9269	0.9907	<u>0.9300</u>	0.9423	0.8641	0.0769
	ETIS-Larib	0.4020	<u>0.9612</u>	0.7102	0.2578	0.2610	0.5991
	Kvasir-SEG	0.4909	0.8829	0.9084	0.3082	0.3302	<u>0.5084</u>

Table 14 Cross-testing results, where Ψ_{net} is trained on ETIS-Larib and tested on CVC-ClinicDB, CVC-ColonDB, ETIS-Larib, and Kvasir-SEG. The highest scores are bolded, and the second place is underlined

Training dataset	Testing datasets	Metrics					
		DSC	ACC	PREC	REC	IoU	loss
ETIS-LaribPolypDB	CVC-ClinicDB	0.5655	0.9348	0.7412	<u>0.5019</u>	0.4023	<u>0.4351</u>
	CVC-ColonDB	0.4327	<u>0.9406</u>	0.7422	0.3372	0.2866	0.5688
	ETIS-Larib	0.8888	0.9967	0.9767	0.8986	0.8000	0.1081
	Kvasir-SEG	0.5841	0.9006	<u>0.9022</u>	0.4419	<u>0.4207</u>	0.4159

indicated, superior performance is achieved in all metrics when Ψ net is trained and tested on CVC-ClinicDB. CVC-ClinicDB achieves the top DSC of 0.9449 compared to Kvasir-SEG which occupies the second place in DSC with 0.7688, while ETIS-Larib and CVC-ColonDB come in the third and fourth place with DSC of 0.7373 and 0.6494, respectively. CVC-ColonDB shows the worst performance in this cross-evaluation process, mainly, because of the high dissimilarity between the original training CVC-ClinicDB dataset and the testing CVC-ColonDB one. Figure 25 indicates some visual samples of cross-evaluating the proposed model.

4.2.2 Kvasir-SEG-based cross-evaluation

Here, the model is trained on Kvasir-SEG, then the testing process is performed on Kvasir-SEG, besides the other three independent polyp datasets, i.e., CVC-ColonDB, CVC-ClinicDB, and ETIS-Larib. The results of cross-testing the proposed Ψ net are shown in Table 12. As indicated, the proposed methodology achieves the highest scores in all metrics when Ψ net is trained and tested on Kvasir-SEG. Kvasir-SEG attains the highest DSC of 0.9045, compared to testing on CVC-ClinicDB which occupies the second place in DSC with a score of 0.8120, while CVC-ColonDB and ETIS-Larib come in the third and the fourth place with DSC of 0.7215 and 0.6433, respectively, via cross-evaluation. ETIS-Larib shows the lowest performance in this cross-evaluation process in terms of DSC, precision, and IoU due to the varying nature of polyps because of their unique shapes,

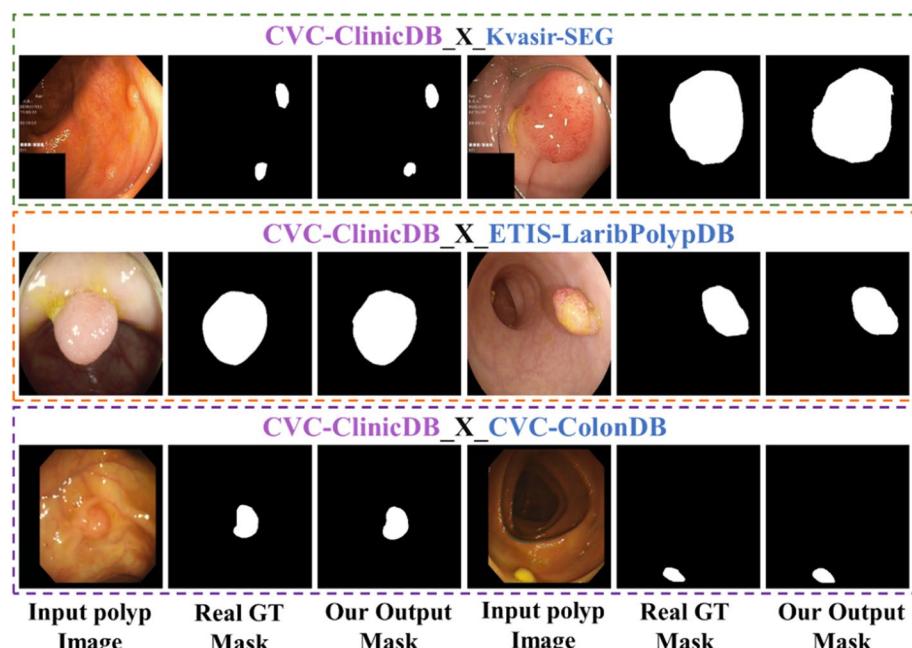


Fig. 25 Visual results of cross-testing the proposed model. The training process is performed on CVC-ClinicDB while testing is performed on Kvasir-SEG, ETIS-Larib, CVC-ColonDB, besides the original dataset employed in the training process

appearance, different colors, sizes, and structures. Figure 26 shows some visual samples of cross-evaluating the proposed model.

4.2.3 CVC-ColonDB-based cross-evaluation

Here, the model is trained on CVC-ColonDB and the testing process is performed on CVC-ColonDB, besides the other three independent polyp datasets, i.e., CVC-ClinicDB, ETIS-Larib, and Kvasir-SEG. Table 13 shows the results of cross-testing the proposed Ψ net. From Table 13, the best high scores are achieved when Ψ net is trained and tested on CVC-ColonDB. CVC-ColonDB achieves the best DSC of 0.9269 compared to testing on CVC-ClinicDB which takes the second DSC place of 0.6075, while Kvasir-SEG and ETIS-Larib come in the third and the fourth place with DSC of 0.4909 and 0.4020, respectively. ETIS-Larib shows the lowest performance in this cross-evaluation process due to the varying nature of CVC-ColonDB polyps compared to ETIS-Larib, such as variation in pixel intensity distribution generated using different colonoscopes. Figure 27 indicates some visual samples of cross-evaluating the proposed model.

4.2.4 ETIS-Larib-based cross-evaluation

Here, the model is trained on ETIS-Larib and the testing process is performed on ETIS-Larib, besides the other three independent polyp datasets, i.e., CVC-ColonDB, Kvasir-SEG, and CVC-ClinicDB. Table 14 shows the results of cross-testing the proposed Ψ net.

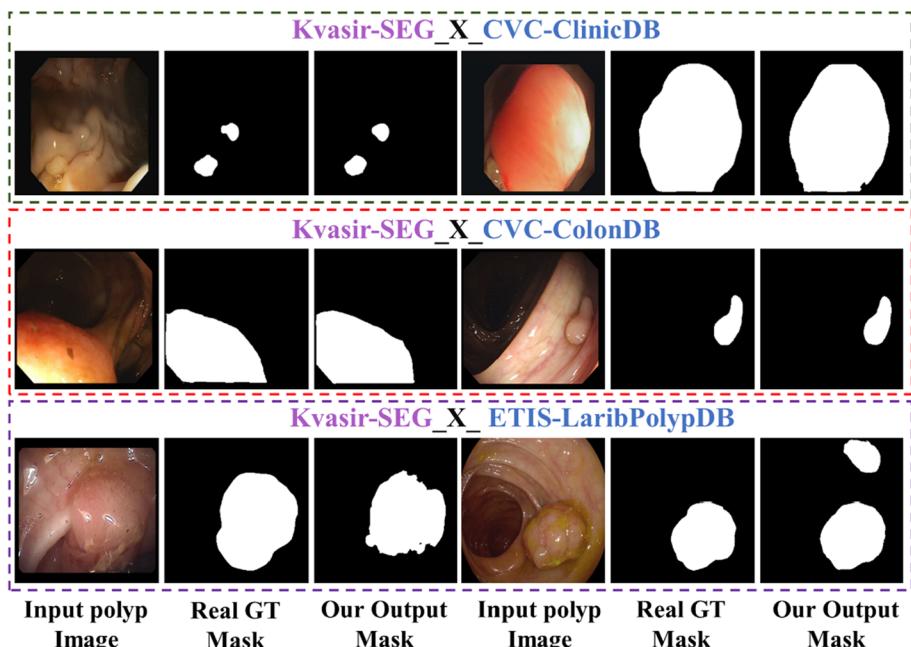


Fig. 26 Visual results of cross-testing the proposed model. The training process is performed on Kvasir-SEG while testing is performed on CVC-ClinicDB, CVC-ColonDB, ETIS-Larib, besides the original dataset employed in the training process

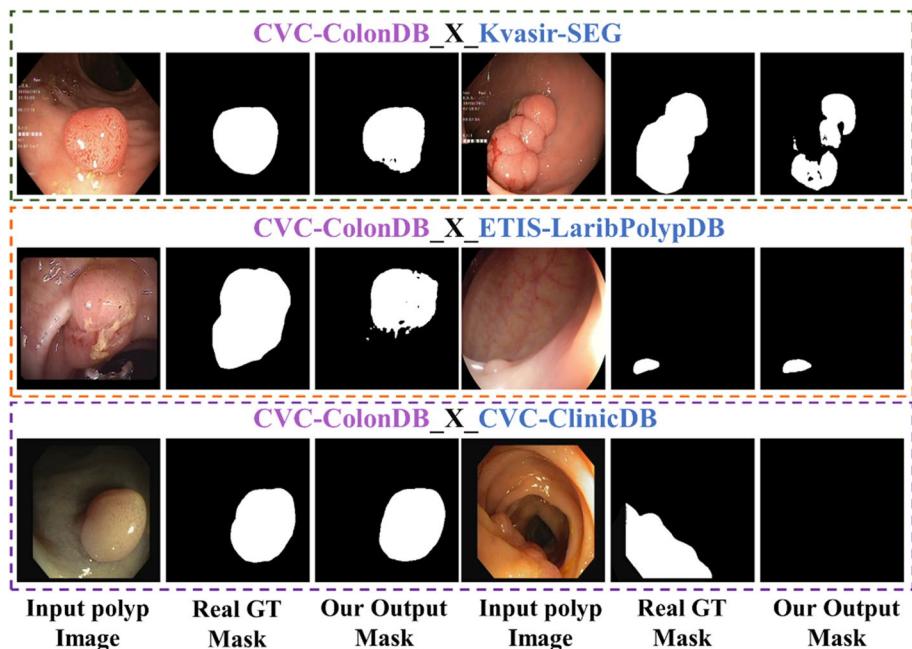


Fig. 27 Visual results of cross-testing the proposed model. The training process is performed on CVC-ColonDB while testing is performed on Kvasir-SEG, ETIS-Larib, CVC-ClinicDB, besides the original dataset employed in the training process

As indicated, Ψ net shows the top performance when it is trained and tested on ETIS-Larib. ETIS-Larib accomplishes the best DSC of 0.8888, compared to Kvasir-SEG which occupies the second place with a DSC of 0.5841, while CVC-ClinicDB and CVC-ColonDB come in the third and fourth places with DSC of 0.5655 and 0.4327, respectively. CVC-ColonDB shows the worst performance in this cross-evaluation process in terms of DSC, IoU, and recall, mainly because of the high dissimilarity between ETIS-Larib and CVC-ColonDB, each dataset has its nature and characteristics. Figure 28 shows some visual samples of cross-evaluating the proposed model.

4.3 Ablation study

To verify the effectiveness of Ψ net, ablation studies are conducted to analyze various elements and settings, including hyper-parameter tuning, loss function, image resolution, and pretrained network.

Hyper-parameters Hyper-parameter tuning is an essential task in deep learning models, which determines the accuracy of the model. Hyper-parameters control the learning process which consequently impacts how well the model performs. An optimal configuration gives the best results whereas suboptimal choices may result in the worst accuracy. Generally, after multiple trials, and error basis evaluation, our hyper-parameters are chosen. However, it is time-consuming to test various combinations without any scientific reason.

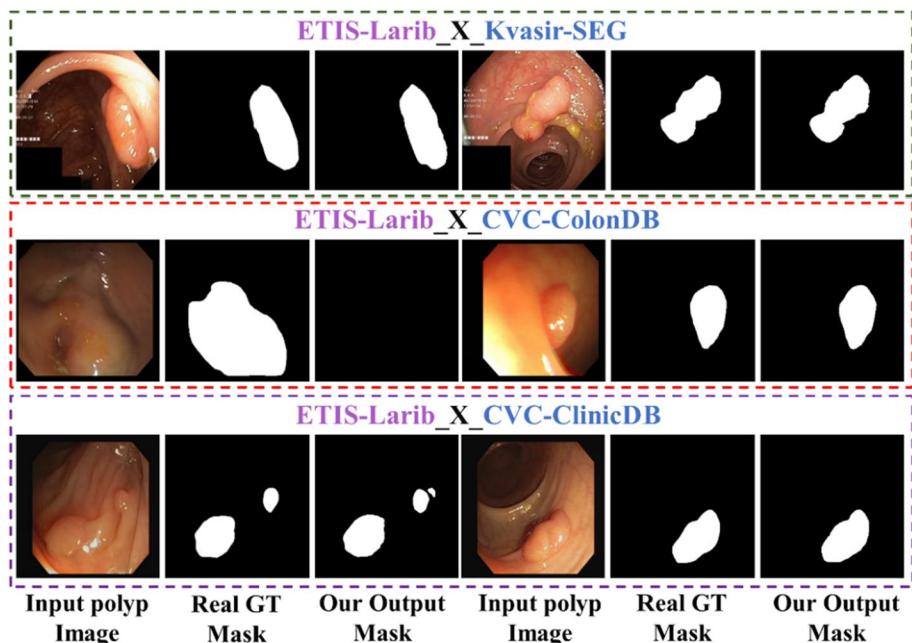


Fig. 28 Visual results of cross-testing the proposed model. The training process is performed on ETIS-Larib while testing is performed on Kvasir-SEG, CVC-ColonDB, CVC-ClinicDB, besides the original dataset employed in the training process

Batch size is defined as the number of examples that are used from the training dataset to estimate the gradient error which influences the dynamics of the learning algorithm. Large batch sizes slow down the learning process but produce more stable models compared to smaller batches. In our experiments, a batch size of 16 was chosen. Using activation functions, like ReLU, in the different modules inside the proposed system helped in resolving the vanishing gradient problem. Adaptive learning rate optimizers, like Adam, show efficient performance and achieve higher accuracy with a learning rate of 0.0001.

Changing the image size greatly affects the accuracy and the training time of the model and its memory space. Hence, it is the main objective to set the best proper image size in the training datasets to extract the essential information such as size, shape, and texture that could contribute to the improvement in the image segmentation accuracy. Employing an image size of 256×256 achieves a good balance between the computation time and the performance. In addition, it is like a standard size in all alternative models that we have employed in our experiments.

The performance of the SE module is controlled by its reduction ratio (r). Increasing the reduction ratio reduces the total number of parameters of Ψ net. For example, by changing r from 8 to 16, the total number of parameters changed from 33,903,845 to 33,849,445. Mostly, setting $r=8$ assures a good balance between complexity and accuracy. Hence, it is used as the default reduction ratio in Ψ net as in Tables 3, 4 and 5, while in Tables 6, 7, 8, 9 and 10, we show the effect of increasing the reduction ratio on performance. As indicated, sometimes, we can get better performance with larger r such as the results of ETIS-LaribPolypDB.

Loss function In highly imbalanced segmentation tasks, small-sized foreground classes in the training process are ignored, which results in low segmentation accuracy. This is called the class imbalance problem and can be alleviated by weighting the loss of the small-sized foreground classes. Loss functions are an important factor in handling this problem. There are two types of class imbalance, i.e., class imbalance at the sample-level and pixel-level. Pixel-level class imbalance occurs when only a few pixels of a sample addressing a particular class are harder to be addressed at the data collection stage compared to other ones. On the other side, the sample-level class imbalance describes the imbalance of classes in a dataset. Like classification tasks, this type of imbalance can be addressed during data collection by including class representatives uniformly.

Existing loss functions for segmentation tasks can be divided into four categories [21, 33]: distribution-based loss, region-based loss, boundary-based loss, and compounded loss. Distribution-based loss functions, such as cross-entropy and focal loss, measure the dissimilarity between two distributions. Region-based loss functions quantify the mismatch or the overlap between two regions, such as dice loss, Tversky loss, focal Tversky loss, and log-cosh dice loss. Boundary-based loss functions measure the distance between two boundaries, such as Euclidean distance or Harsdorf distance. Compounded loss functions are defined as the combinations among the distribution-, region-, and boundary-based loss functions (combo loss). Most of the loss functions are based on cross-entropy and dice loss functions. However, the objects in medical images, such as nuclei and polyps often occupy a small region in the image. The cross-entropy loss and some others are not optimal for such tasks. Most SOTA methods employ the dice coefficient loss function in their experiments. Hence, we follow their steps for the sake of fair comparisons. However, in Table 15, the effect of different loss functions is shown on ISIC-2018. As shown, the scores are very similar and close, but focal Tversky loss shows the best performance on ISIC-2018. ISIC-2018 is a heavily imbalanced dataset, hence it is the most one affected by changing the loss function.

The impact of pretrained network Here, we replaced the adopted VGG-19 in the first encoder of NET1 with other common pretrained networks, such as ResNet50, and DenseNet121. Both ResNet50 and DenseNet121 can provide a lower number of parameters but at expense of the performance, see Table 16 for a computitative comparison of

Table 15 Performance evaluation of different types of segmentation loss functions on ISIC-2018. The highest scores are bolded, and the second place is underlined

Loss Function	Evaluation Metrics					
	F1-score	ACC	PREC	REC	IoU	SPEC
Dice loss	0.9305	0.9639	0.9484	0.9132	0.8700	0.9822
Dice loss squared	0.9303	0.9642	0.9595	0.9028	0.8696	0.9863
Generalized dice loss	0.9287	0.9638	<u>0.9684</u>	0.8922	0.8669	<u>0.9895</u>
Dice focal loss	0.9254	0.9623	0.9703	0.8845	0.8612	0.9903
Combo loss	0.9322	<u>0.9652</u>	0.9601	0.9058	0.8730	0.9865
Focal tversky loss	0.9331	0.9653	0.9520	<u>0.9150</u>	0.8746	0.9834
Tversky loss	<u>0.9323</u>	0.9647	0.9454	0.9195	<u>0.8731</u>	0.9809
Log-cosh dice loss	0.9271	0.9628	0.9622	0.8945	0.8641	0.9874
Focal loss	0.9222	0.9604	0.9595	0.8877	0.8556	0.9865

Table 16 Ablation study on the influence of pretrained networks in Ψ_{net} on CVC-ColonDB and ETIS-LaribPolyPDB

Dataset	Pretrained network	Total Ψ_{net} parameters	Non-trainable parameters	DSC	IoU	REC	PREC	ACC
CVC-ColonDB	ResNet50	23,310,085	42,496	0.9160	0.8457	0.9505	0.9239	0.9907
	DenseNet121	22,390,853	7,050,688	0.5843	0.4154	0.6865	0.8523	0.9712
	VGG-19	33,849,445	11,904	0.9299	0.8693	0.9359	0.9295	0.9903
ETIS-LaribPolyPDB	ResNet50	23,310,085	42,496	0.8841	0.7930	0.9150	0.9708	0.9969
	DenseNet121	22,390,853	7,050,688	0.3012	0.1777	0.0428	0.9934	0.9758
	VGG-19	33,849,445	11,904	0.9343	0.8768	0.9118	0.9771	0.9970

CVC-ColonDB and ETIS-LaribPolypDB. These two datasets are the most challenging ones. They have polyps with a variety of sizes, shapes, textures, and characteristics as depicted in Figs. 19, 20, 21 and 22. Some of them have the same texture as the colon and grow horizontally, leading to polyp misdetections. Hence, extracting the salient features by the pretrained network in the first encoder is a challenging task.

5 Conclusion

This paper introduces a novel encoder-decoder-based architecture, dubbed Ψ net, to semantically segment medical images. A pre-trained network is employed in the proposed encoder to increase the model's capability of learning long-range dependencies and capturing global contextual representation effectively. In addition, we have increased the representation of salient features by weighing every feature map with a squeeze-and-excitation block. Moreover, we have added ASPP for dense large-scale feature extraction by capturing global multiscale contextual information. Hence, the proposed model combines the targeted semantic information at different levels. We validated the effectiveness of the proposed Ψ net via extensive experiments on different segmentation tasks with different modalities, such as colonoscopy, dermoscopy, and microscopy. In these experiments, the proposed Ψ net achieves superior performance compared to SOTA models, such as U-Net, ResUNet, ResUNet++, UNet++, BCDU-Net, MCGU-Net, FRCU-Net, Attention Deeplabv3p, DDANet, ColonSegNet, and TMD-Unet. In all employed datasets, our model produced the absolute best DSC results despite the challenges of having different shapes, types, and sizes, ranging from tiny to enormous, in polyps, lesions, and nuclei. In addition, for testing the generalizability of the proposed model, a cross-evaluation is performed on different datasets which proves good performance, especially when the training and testing datasets share the same nature. Hence, the proposed Ψ net could be a new baseline in medical image segmentation.

Model limitations and future work Despite the effectiveness of the proposed Ψ net, it comprises three parallel multi-scale branches, which leads to a network of around 33 M parameters compared to the traditional U-Net of 8 M. This number of parameters makes the model more complex and slower in training. Hence, in the future, we intend to employ attentive and residual mechanisms that may help to reduce the branches' complexity. In addition, we believe that increasing the dataset size and implementing additional augmentation approaches will boost the model's performance even further. Moreover, the application of Ψ net should not be confined to medical segmentation but might be extended to natural image segmentation and other pixel-wise classification tasks. Furthermore, we will seek to extend the proposed Ψ net to include volumetric segmentation tasks.

Funding Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB).

Data availability We have cited all employed datasets. Data will be available in reasonable request.

Declaration of interests

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.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M (2016) {TensorFlow}: a system for {Large-Scale} machine learning. In: 12th USENIX symposium on operating systems design and implementation (OSDI 16), pp 265–283
2. Akbari M, Mohrekesh M, Nasr-Esfahani E, Soroushmehr SR, Karimi N, Samavi S, Najarian K (2018, July) Polyp segmentation in colonoscopy images using fully convolutional network. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, pp 69–72
3. Al-Masni MA, Al-Antari MA, Choi MT, Han SM, Kim TS (2018) Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks. *Comput Methods Prog Biomed* 162:221–231
4. Alom MZ, Yakopcic C, Hasan M, Taha TM, Asari VK (2019) Recurrent residual U-Net for medical image segmentation. *J Med Imaging* 6(1):014006
5. Anwar SM, Majid M, Qayum A, Awais M, Alnowami M, Khan MK (2018) Medical image analysis using convolutional neural networks: a review. *J Med Syst* 42(11):1–13
6. Asadi-Aghbolaghi M, Azad R, Fathy M, Escalera S (2020) Multi-level context gating of embedded collective knowledge for medical image segmentation. *arXiv preprint arXiv:2003.05056*
7. Azad R, Asadi-Aghbolaghi M, Fathy M, Escalera S (2019) Bi-directional ConvLSTM U-Net with densely connected convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp 0–0
8. Azad R, Asadi-Aghbolaghi M, Fathy M, Escalera S (2020, August) Attention deeplabv3+: Multi-level context attention mechanism for skin lesion segmentation. In: European Conference on Computer Vision, Springer, Cham, pp 251–266
9. Azad R, Bozorgpour A, Asadi-Aghbolaghi M, Merhof D, Escalera S (2021) Deep frequency re-calibration U-Net for medical image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 3274–3283
10. Bernal J, Sánchez J, Vilarino F (2012) Towards automatic polyp detection with a polyp appearance model. *Pattern Recogn* 45(9):3166–3182
11. Bernal J, Sánchez FJ, Fernández-Esparrach G, Gil D, Rodríguez C, Vilariño F (2015) WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Comput Med Imaging Graph* 43:99–111. <https://doi.org/10.1016/j.compmedimag.2015.02.007>
12. Caicedo JC, Goodman A, Karhohs KW, Cimini BA, Ackerman J, Haghghi M, Heng C, Becker T, Doan M, McQuin C, Rohrbach M (2019) Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. *Nat Methods* 16(12):1247–1253 Available: <https://www.kaggle.com/c/data-science-bowl-2018>
13. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille 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
14. Chen L, Bentley P, Mori K, Misawa K, Fujiwara M, Rueckert D (2018) DRINet for medical image segmentation. *IEEE Trans Med Imaging* 37(11):2453–2462
15. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp 801–818
16. Chen LC, Papandreou G, Schroff F, Adam H (2019) Rethinking atrous convolution for semantic image segmentation. *arXiv* 2017. *arXiv preprint arXiv:1706.05587*, 2
17. Chollet F (2018) Keras: The python deep learning library. *Astrophysics source code library ascl-1806*
18. Codella NC, Gutman D, Celebi ME, Helba B, Marchetti MA, Dusza SW, Kalloo A, Liopyris K, Mishra N, Kittler H, Halpern A (2018, April) Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international

- skin imaging collaboration (isic). In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018), IEEE, pp 168–172
19. Codella N, Rotemberg V, Tschandl P, Celebi ME, Dusza S, Gutman D, Helba B, Kalloo A, Liopyris K, Marchetti M, Kittler H (2019) Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368
 20. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2020) Imagenet: A large-scale hierarchical image database, 2009. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 248–255
 21. El Jurdi R, Petitjean C, Honeine P, Cheplygina V, Abdallah F (2021) High-level prior-based loss functions for medical image segmentation: A survey. *Comput Vis Image Underst* 210:103248
 22. Erol T, Sarikaya D (2022) PlutoNet: An efficient polyp segmentation network. arXiv preprint arXiv:2204.03652
 23. Eu CY, Tang TB, Lin CH, Lee LH, Lu CK (2021) Automatic polyp segmentation in colonoscopy images using a modified deep convolutional encoder-decoder architecture. *Sensors* 21(16):5630
 24. Fan DP, Ji GP, Zhou T, Chen G, Fu H, Shen J, Shao L (2020, October) PraNet: Parallel reverse attention network for polyp segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer, Cham, pp 263–273
 25. Fang Y, Chen C, Yuan Y, Tong KY (2019, October) Selective feature aggregation network with area-boundary constraints for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, pp 302–310
 26. Fu Y, Lei Y, Wang T, Curran WJ, Liu T, Yang X (2021) A review of deep learning-based methods for medical image multi-organ segmentation. *Phys Med* 85:107–122
 27. Gu K, Xia Z, Qiao J, Lin W (2019) Deep dual-channel neural network for image-based smoke detection. *IEEE Trans Multimed* 22(2):311–323
 28. Gu K, Zhang Y, Qiao J (2020) Ensemble meta-learning for few-shot soot density recognition. *IEEE Trans Ind Inform* 17(3):2261–2270
 29. Gu K, Liu H, Xia Z, Qiao J, Lin W, Thalmann D (2021) PM_{2.5} monitoring: use information abundance measurement and wide and deep learning. *IEEE Trans Neural Netw Learn Syst* 32(10):4278–4290
 30. He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans Pattern Anal Mach Intell* 37(9):1904–1916
 31. 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
 32. Ibtehaz N, Rahman MS (2020) MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Netw* 121:74–87
 33. Jadon S (2020, October) A survey of loss functions for semantic segmentation. In: 2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB). IEEE, pp 1–7
 34. Jha D, Smedsrød PH, Riegler MA, Johansen D, De Lange T, Halvorsen P, Johansen HD (2019, December) Resunet++: An advanced architecture for medical image segmentation. In: 2019 IEEE International Symposium on Multimedia (ISM), IEEE, pp 225–2255
 35. Jha D, Riegler MA, Johansen D, Halvorsen P, Johansen HD (2020, July) 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
 36. Jha D, Smedsrød PH, Riegler MA, Halvorsen P, Lange TD, Johansen D, Johansen HD (2020, January) Kvadir-seg: A segmented polyp dataset. In: International Conference on Multimedia Modeling, Springer, Cham, pp. 451–462. Available: <https://datasets.simula.no/kvadir-seg/>
 37. Jha D, Ali S, Tomar NK, Johansen HD, Johansen D, Rittscher J, Riegler MA, Halvorsen P (2021) Real-time polyp detection, localization and segmentation in colonoscopy using deep learning. *IEEE Access* 9:40496–40510
 38. Jha D, Tomar NK, Ali S, Riegler MA, Johansen HD, Johansen D, de Lange T, Halvorsen P (2021, June) Nanonet: Real-time polyp segmentation in video capsule endoscopy and colonoscopy. In: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS). IEEE, pp 37–43
 39. Jha D, Smedsrød PH, Johansen D, de Lange T, Johansen HD, Halvorsen P, Riegler MA (2021) A comprehensive study on colorectal polyp segmentation with ResUNet++, conditional random field and test-time augmentation. *IEEE J Biomed Health Inform* 25(6):2029–2040
 40. Li X, Chen D (2022) A survey on deep learning-based panoptic segmentation. *Digit Signal Process* 120:103283
 41. Liu X, Song L, Liu S, Zhang Y (2021) A review of deep-learning-based medical image segmentation methods. *Sustainability* 13(3):1224
 42. 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
 43. Minaee S, Boykov YY, Porikli F, Plaza AJ, Kehtarnavaz N, Terzopoulos D (2021) Image segmentation using deep learning: A survey. *IEEE Trans Pattern Anal Mach Intell* 44(7):3523–3542

44. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla NY, Kainz B, Glocker B (2018) Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999
45. Park KB, Lee JY (2022) SwinE-Net: hybrid deep learning approach to novel polyp segmentation using convolutional neural network and Swin Transformer. *J Comput Des Eng* 9(2):616–632
46. Patel D, Shah Y, Thakkar N, Shah K, Shah M (2020) Implementation of artificial intelligence techniques for cancer detection. *Augment Hum Res* 5(1):1–10
47. Punn NS, Agarwal S (2022) Modality specific U-Net variants for biomedical image segmentation: a survey. *Artif Intell Rev* 55:1–45
48. Razzaq MI, Naz S, Zaib A (2018) Deep learning for medical image processing: Overview, challenges and the future. *Classification in BioApps*, pp 323–350
49. Ronneberger O, Fischer P, Brox T (2015, October) U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, pp 234–241
50. Silva J, Histace A, Romain O, Dray X, Granado B (2014) Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *Int J Comput Assist Radiol Surg* 9(2):283–293
51. Simonyan K, Zisserman A (2018) Very deep convolutional networks for large-scale image recognition. *Karen. Am J Health Pharm* 75:398–406
52. Sudre CH, Li W, Vercauteren T, Ourselin S, Jorge Cardoso M (2017) Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3, Springer International Publishing 240–248
53. Suganyadevi S, Seethalakshmi V, Balasamy K (2022) A review on deep learning in medical image analysis. *Int J Multimed Inf Retr* 11(1):19–38
54. Tajbakhsh N, Gurudu SR, Liang J (2015) Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Trans Med Imaging* 35(2):630–644
55. Tomar NK, Jha D, Ali S, Johansen HD, Johansen D, Riegler MA, Halvorsen P (2021, January) DDANet: Dual decoder attention network for automatic polyp segmentation. In: International Conference on Pattern Recognition. Springer, Cham, pp 307–314
56. Tran ST, Cheng CH, Nguyen TT, Le MH, Liu DG (2021) TMD-Unet: Triple-Unet with multi-scale input features and dense skip connection for medical image segmentation. *Healthcare* 9(1):54 MDPI
57. Tschandl P, Rosendahl C, Kittler H (2018) The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data* 5(1):1–9
58. Verma R, Kumar N, Patil A, Kurian NC, Rane S, Graham S, Vu QD, Zwager M, Raza SEA, Rajpoot N, Wu X (2021) MoNuSAC2020: A multi-organ nuclei segmentation and classification challenge. *IEEE Trans Med Imaging* 40(12):3413–3423
59. Wang J, Sun K, Cheng T, Jiang B, Deng C, Zhao Y, Liu D, Mu Y, Tan M, Wang X, Liu W (2020) Deep high-resolution representation learning for visual recognition. *IEEE Trans Pattern Anal Mach Intell* 43(10):3349–3364
60. Wei J, Hu Y, Zhang R, Li Z, Zhou SK, Cui S (2021, September) Shallow attention network for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, pp. 699–708
61. Wu Y, Lin L, Wang J, Wu S (2020) Application of semantic segmentation based on convolutional neural network in medical images. *Sheng Wu Yi Xue Gong Cheng Xue Za Zhi= Journal of Biomedical Engineering= Shengwu Yixue Gongchengxue Zazhi* 37(3):533–540
62. Xing Y, Zhong L, Zhong X (2020) An encoder-decoder network-based FCN architecture for semantic segmentation. *Wirel Commun Mob Comput* 2020:1–9
63. Zahangir Alom M, Hasan M, Yakopcic C, Taha TM, Asari VK (2018) Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation. arXiv e-prints, pp arXiv-1802
64. Zhang L, Dolwani S, Ye X (2017, July) Automated polyp segmentation in colonoscopy frames using fully convolutional neural network and textons. In: Annual Conference on Medical Image Understanding and Analysis. Springer, Cham, pp 707–717
65. Zhang Z, Liu Q, Wang Y (2018) Road extraction by deep residual u-net. *IEEE Geosci Remote Sens Lett* 15(5):749–753
66. Zhao X, Zhang L, Lu H (2021, September) Automatic polyp segmentation via multi-scale subtraction network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, pp 120–130

67. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J (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 International Publishing, 3–11 2018
68. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J (2019) Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans Med Imaging* 39(6):1856–1867
69. Zunair H, Hamza AB (2021) Sharp U-Net: Depthwise convolutional network for biomedical image segmentation. *Comput Biol Med* 136:104699

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.