

Article

XCC-Net: An X-Shaped Collective Convolution Network Architecture for Medical Image Segmentation

Anass Garbaz ^{1,*} , Yassine Oukdach ¹, Said Charfi ¹ , Mohamed El Ansari ² , Lahcen Koutti ¹,
Mustapha Hedabou ³, Mustapha Oujoura ⁴  and Abdel Motalib Lagsoun ⁴ 

¹ Laboratory of Computer Systems and Vision, Department of Computer Science, Faculty of Sciences, Ibn Zohr University, Agadir 80000, Morocco; yassine.oukdach@edu.uiz.ac.ma (Y.O.); s.charfi@uiz.ac.ma (S.C.); l.koutti@uiz.ac.ma (L.K.)

² Informatics and Applications Laboratory, Department of Computer Science Faculty of Sciences, My Ismail University, Meknes 50000, Morocco; moh.elansari@umi.ac.ma

³ College of Computing, Mohammed VI Polytechnic University (UM6P), Ben Guerir 43150, Morocco; mustapha.hedabou@um6p.ma

⁴ Mathematics, Informatics & Communication Systems Laboratory (MISCOM), National School of Applied Sciences of Safi, Cadi Ayyad University, Marrakech 40000, Morocco; mustapha.oujaoura@uca.ac.ma (M.O.); a.lagsoun.ced@uca.ac.ma (A.M.L.)

* Correspondence: a.garbaz@uiz.ac.ma

Abstract

Encoder–decoder models are widely used for pixel-level segmentation due to their ability to capture and combine multiscale features. However, skip connections between the encoder and decoder often require cropping to mitigate border pixel loss during convolutions, which can introduce inefficiencies and limit performance. This study explores the potential of modifying these connections by removing direct encoder-to-decoder links to enhance segmentation accuracy. We propose a novel architecture, termed XCC-Net, which features two context-capturing pathways and two symmetric pathways for enlargement. These pathways are interconnected via channels, enabling automated detection of structures with varied shapes. The XCC-Net's X-shaped architecture links skip connections exclusively between encoder-to-encoder and decoder-to-decoder, omitting direct encoder-to-decoder feature transfers to potentially improve performance. The XCC-Net model was evaluated on multiple medical imaging datasets, including wireless capsule endoscopy (WCE), colonoscopy, and dermoscopy images. Experimental results showed that XCC-Net outperformed state-of-the-art segmentation models, achieving dice coefficients of 91.70%, 89.26%, 87.15%, and 79.07% on the MICCAI 2017 (Red Lesion), PH2, CVC-ClinicDB, and ISIC 2017 datasets, respectively. XCC-Net's X-shaped architecture, with its unique skip connections, demonstrates improved segmentation performance across various medical imaging tasks.

Keywords: XCC-Net; feature fusion; medical image segmentation; skip connections



Academic Editor: Guoqing Chao

Received: 8 November 2025

Revised: 12 December 2025

Accepted: 23 December 2025

Published: 25 December 2025

Copyright: © 2025 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Deep learning has recently achieved significant breakthroughs in semantic, pixel-level annotation within the field of medical imaging [1,2].

Particularly, fully convolutional networks (FCNs) [3] have risen to prominence in the segmentation of medical images. Among the leading contemporary models stands U-Net [4]. Following the introduction of the initial U-Net framework, numerous adaptations and enhancements have been suggested.

Additionally, encoder–decoder models often employ skip connections linking corresponding encoder and decoder layers. While these connections facilitate information transfer and spatial detail recovery, they can also complicate model interpretability by obscuring the decision-making process. The direct links between the encoder and decoder layers may obscure the logic behind the model’s predictions, making it more challenging to explain why specific regions are segmented in a particular way.

In this paper, we introduce XCC-Net, an X-shaped architecture in which skip connections are established only within the encoder paths and within the decoder paths, without direct connections between encoders and decoders. The network comprises two distinct encoder subnetworks: the X-Separable Encoder (XSE) and the Multi-Channel Separable Encoder (MCSE). Outcomes from both encoder networks feed into a Global-Feature Ensembling (GFE) module.

These outcomes are subsequently relayed to two corresponding upsampling decoders, namely the X-Separable Decoder (XSD) and the Multi-Channel Separable Decoder (MCSD). The XSE and MCSE encoders are interconnected through shortcut connections. These connections are made among layers of identical hierarchical status. The skip connections traverse a Powered-Feature Engagement (PFE) module.

From a clinical perspective, skin and gastrointestinal cancers are among the fastest-growing malignancies worldwide. For example, approximately 104,930 cases of skin cancer and 348,840 cases of gastrointestinal cancer were diagnosed in the United States in 2023 [5].

Skin cancer [6] is characterized by the uncontrolled proliferation of abnormal cells within the epidermis resulting from unrepaired DNA damage that induces mutations. Nonetheless, melanoma [7] stands as the most lethal variant of skin cancer, originating from melanocytes, the cells responsible for melanin production. Dermatoscopy [8] serves as a non-invasive imaging technique utilized in dermatology for scrutinizing skin lesions. However, current clinical protocols rely on manual visual examination for early melanoma detection, a process that is subjective, time-intensive, and lacks reproducibility.

Gastrointestinal (GI) disorders [9] encompass a broad spectrum of ailments that exhibit considerable variability in severity and potential hazards or complications. While some GI conditions are controllable and may only cause mild discomfort or inconvenience, others carry substantial health risks. A common concern observed in both hospital emergency departments and the broader healthcare setting is GI bleeding [10]. Symptoms such as hematemesis, melena, or hematochezia indicate acute or overt GI hemorrhage. In contrast, cryptic GI bleeding primarily originates from the small intestine and involves recurrent episodes. The source remains unidentified even after an upper colonoscopy evaluation [11].

Prior to the development of Wireless Capsule Endoscopy (WCE) [12], there was no technology available to inspect hemorrhages in the small intestine. This pill-shaped device incorporates an optical crown, light source, camera, power supply, and RF transmitter, and measures 26 mm × 11 mm. The individual is required to swallow a WCE, which broadcasts between 50,000 and 60,000 visual frames over an eight-hour period, averaging 2 to 4 frames per second.

Consequently, manually inspecting WCE frames to identify abnormalities like GI bleeding demands a significant amount of time. Furthermore, WCE encounters various challenges that compromise segmentation accuracy. These include inadequate contrast, obscured surroundings, and variations in lesion appearance and shade.

In this paper, we present the XCC-Net to deal with skin and gastro-intestinal tract abnormalities. The datasets used for such a task are illustrated in [13–16].

2. Related Works

2.1. Architectures

Convolutional networks are robust visual models capable of generating hierarchical features. Recent research [3] demonstrated that these networks, trained end-to-end, pixel by pixel, surpass the state-of-the-art in semantic segmentation.

Building upon the FCN framework, subsequent research, as detailed in [4], introduced a more sophisticated architecture known as U-Net. Additionally, it incorporates skip connections. The inclusion of skip connections maintains spatial details and helps the network restore intricate features during upsampling.

Zhou et al. [17] proposed UNet++. It tackles the challenge of unknown network depth by employing an efficient ensemble of U-Nets with different depths. These networks share an encoder and learn simultaneously through deep supervision.

A novel attention gate model tailored for medical imaging was presented by Oktay et al. [18]. It autonomously adapts to concentrate on target structures of diverse shapes and sizes. In the paper by Iglovikov et al. [19], it was illustrated how enhancing the U-Net-style architecture could be achieved through the integration of a pre-trained encoder. Meanwhile, the Diakogiannis et al. model [20] comprises a novel deep learning structure named ResUNet-a, along with a unique loss function derived from the Dice loss.

ResUNet-a employs a UNet framework, coupled with residual connections, atrous convolutions, pyramid scene parsing pooling, and multi-tasking inference. A novel 2D attention residual U-Net architecture was developed by Lafraxo et al. [21]. It integrates both the attention mechanism and residual units into U-Net to improve polyp and bleeding segmentation performance.

Meanwhile, Chen et al. [22] proposed TransUNet, a hybrid model combining Transformers and U-Net, as a promising solution for medical image segmentation. On one side, the Transformer analyzes tokenized image patches obtained from a feature map generated by a convolutional neural network to capture global contexts. Similarly, Cao et al. [23] proposed Swin-Unet, a Unet-like model based solely on Transformers for medical image segmentation. Tokenized image patches are fed into the Transformer-based U-shaped Encoder–Decoder architecture. Alom et al. [24] introduced Recurrent Residual Convolutional Neural Network models based on U-Net, denoted as RU-Net and R2U-Net respectively.

SegNet [25] is unique in that its decoder uses pooling indices that were computed during the max-pooling step of the corresponding encoder to carry out non-linear upsampling.

2.2. Gastrointestinal Diseases Detection

Following the introduction of WCE, numerous algorithms have been proposed to alleviate the inherent limitations of WCE [26–31].

Early approaches were based on handcrafted and machine learning methodologies; Ghosh et al. [32] adopted an approach where a block surrounding each pixel is selected to extract local statistical features. Caroppo et al. [33] proposed transferring knowledge from natural images using three pre-trained models from the ImageNet dataset. The extracted features are subsequently reduced and passed through a feature fusion algorithm. A streamlined U-Net with reduced encoder–decoder pairs was proposed by Kanakatte et al. [34] for the detection and segmentation of both bleeding and red lesions in endoscopy data.

Weakly supervised techniques, such as those based on class activation maps [35], have been utilized to accomplish bleeding segmentation with minimal annotation efforts in WCE images. Zhang et al. [36] devised a multi-stage architecture incorporating attention blocks. This architecture addresses the segmentation of small bleeding areas. Hajabdollahi et al. [37]

proposed a low-complexity CNN structure for detecting bleeding zones. This structure accepts a single patch as input and outputs a segmented patch of identical size.

Jain et al. [38] developed a hybrid model comprising a modified SegNet model along with GradCAM++ for detecting anomalies in WCE images. Xing et al. [39] introduced a two-branch Attention Guided Deformation Network for classifying WCE images. In the initial stage, attention maps are employed to guide the amplification of lesion regions in the input images.

2.3. Skin Lesion Detection

Numerous researchers have investigated the potential of deep learning models in skin lesion segmentation [40].

Öztürk et al. [41] presented an enhanced FCN architecture tailored for segmenting full-resolution skin lesion images without the need for any pre- or post-processing. Xie et al. [42] proposed a high-resolution feature block comprising three branches: the main branch, spatial attention branch, and channel-wise attention branch. The main branch utilizes high-resolution feature maps to capture spatial details around boundaries.

Li et al. [43] presented a novel dense deconvolutional network for segmenting skin lesions, leveraging residual learning. This network architecture comprises dense deconvolutional layers, chained residual pooling, and hierarchical supervision. Lei et al. [44] introduced a fresh and efficient generative adversarial network. This network architecture integrates two modules: a segmentation module utilizing skip connections and dense convolution U-Net, along with a dual discrimination module. Wu et al. [45] proposed a CNN incorporating an innovative and effective adaptive dual attention module for automated segmentation of skin lesions. This step is crucial yet challenging for the advancement of computer-assisted diagnosis systems for skin diseases.

To streamline the network and minimize the number of parameters, Hasan et al. [46] substituted depth-wise separable convolution for standard convolution. This modification enables the projection of learned discriminating features onto the pixel space at various stages of the encoder. Zhu et al. [47] introduced an effective double-spatial-shift module aimed at improving the vanilla multilayer perceptron. This improvement facilitates communication among distinct spatial locations through double spatial shifts.

Many encoder–decoder models use skip connections to convey spatial information and low-level features from the encoder to the decoder. This facilitates the generation of more accurate segmentation maps. However, these direct skip connections frequently require cropping to compensate for border pixel loss caused by convolutions.

This necessity can introduce architectural inefficiencies, increase computational complexity, and diminish valuable contextual information. In this work, we explored an alternative strategy to overcome these challenges by eliminating direct encoder-to-decoder connections.

Instead, our approach focuses on redefining the flow of information within the network to enhance performance. By removing these direct links, the proposed architecture aims to reduce the limitations associated with traditional skip connections while improving feature fusion and representation learning. This investigation offers fresh perspectives on advancing segmentation accuracy across various medical imaging applications.

3. Methods

In this section, we explored the XCC-Net architecture. We clarified the information flow within each subnetwork and module. This architecture notably eliminates the need for skip connections between an encoder and a decoder.

3.1. XCC-Net

XCC-Net comprises two symmetric paths corresponding to two encoding paths, resulting in a structure akin to X-shaped, as illustrated in Figure 1. The contracting path runs concurrently across two encoding subnetworks: the X-Separable encoder (XSE) and the multi-channel separable encoder (MCSE).

Skip connections pass through a Powered-Feature Engagement (PFE) module. The outputs from both encoder subnetworks are element-wise added together and then fed into a bottleneck module called global-feature ensembling (GFE). In the last phase, the outputs are directed towards the upsampling decoders, namely the X-Separable decoder (XSD) subnetwork and the multi-channel separable decoder (MCSD) subnetwork.

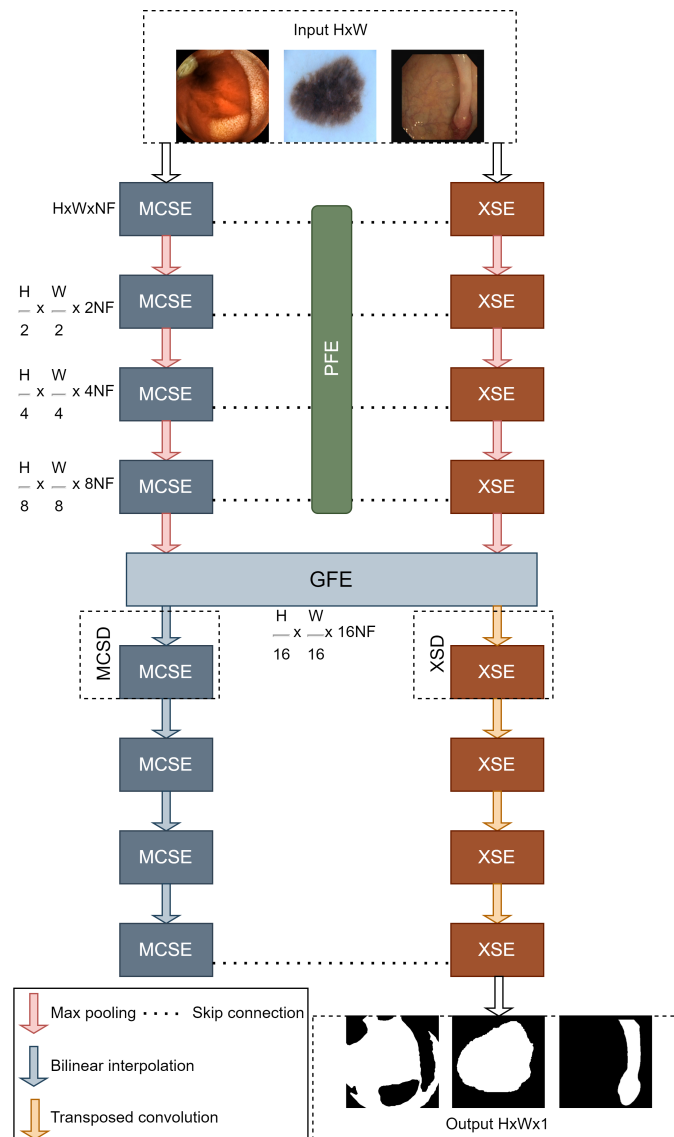


Figure 1. An overview of the proposed XCC-Net architecture illustrating its symmetric X-shaped design. The network includes two parallel encoding subnetworks, XSE and MCSE, connected through the PFE module. The outputs of these encoders are passed to the GFE bottleneck. The decoding stage consists of two corresponding subnetworks, XSD and MCSD, which produce the final segmentation output.

The XSD employs transposed convolution. Meanwhile, MCSD employs bilinear interpolation. The number of filters, denoted as NF, falls within the range $\{64, 128, 256, 512, 1024\}$ for both subnetworks. Additionally, the MCSE subnetwork consists of a varying number of

separated channels, represented as N , which ranges from $\{8, 16, 32\}$. These separated channels are then reconfigured into an ensemble format, resulting in the sequence $\{32, 16, 8\}$ in the MCSD.

3.2. X-Separable Encoder Subnetwork

The XSE block, as illustrated in Figure 2, comprises a repetitive structure that processes the feature map $X \in \mathbb{R}^{H \times W \times C}$, where H , W , and C represent the height, width, and number of channels of the input feature map, respectively.

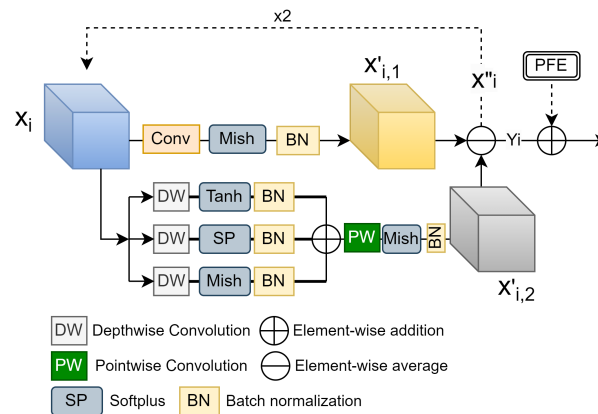


Figure 2. An overview of the XSE architecture. The input passes through two sub-blocks: standard and three depthwise convolutions with different activations. Their outputs are summed and averaged element-wise. This process is repeated, and the final result is summed with the PFE module output before being forwarded to the other encoding path.

In the first transformation within the XSE block, the feature map undergoes a three successive depthwise separable convolutions. Each depthwise convolution operates by applying a single convolutional filter per input channel. This method is effectively described by the mathematical expression:

$$DW_{k,l,n}(X) = \sum_{i,j} K_{i,j,n} \times X_{k+i-1,l+j-1,n} \quad (1)$$

This convolution operation involves shifting the filter $K_{i,j,n}$ across the spatial dimensions of X . The concept is fundamental in Scattering Convolution Networks [48]. Additionally, distinct activation functions are applied to the outputs of the three depthwise convolutions. This promotes inter-channel communication, preserves local feature representations, and expands the local receptive field. The three activation functions employed are:

$$\text{Tanh}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (2)$$

$$\text{Softplus}(x) = \log(e^x + 1) \quad (3)$$

$$\text{Mish}(x) = x \times \text{Tanh}(\text{Softplus}(x)) \quad (4)$$

Feature map variation is introduced through element-wise multiplication with the input. This is followed by summation and pointwise convolution to capture cross-channel interactions. Further stabilization is achieved via Mish activation.

$$X'_{i,2} = \text{Mish}(DW_{k,l,n}(X_i)) + \text{Softplus}(DW_{k,l,n}(X_i)) + \text{Tanh}(DW_{k,l,n}(X_i)) \quad (5)$$

In parallel, a standard convolution operation is applied to the same input. This is followed by a Mish activation function to maintain the traditional flow of feature generation.

The resulting feature map is denoted as $X'_{i,1}$. Subsequently, element-wise averaging (as per Equation (6)) is performed on $X'_{i,1}$ and $X'_{i,2}$ to create a robust feature map. This incorporates both outputs without disregarding any potentially important information from either block.

$$X''_i = \frac{\sum_{m=1}^2 X'_m}{2} \quad (6)$$

The outlined procedure is iterated twice. This amplifies the module's capacity to capture both local and positional information. The XSD block, a fusion of transposed convolution for trainable upsampling and XSE blocks, is then formed.

3.3. Multi-Channel Separable Encoder Subnetwork

The presented MCSE architecture is shown in Figure 3. MCSE is designed to progressively fuse features, capturing both semantic and structural details of lesions from feature representations.

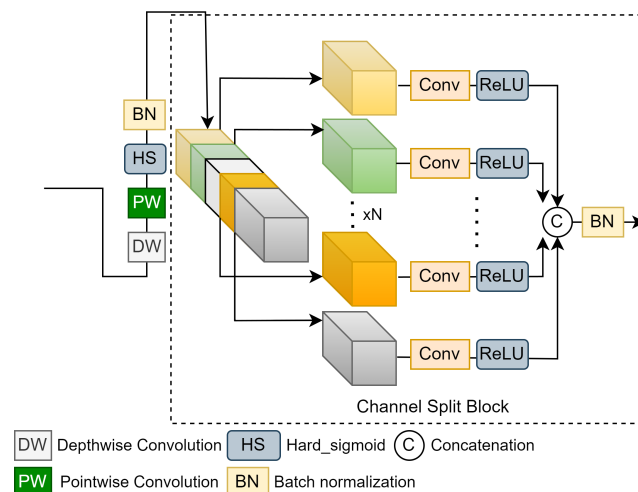


Figure 3. An overview of the MCSE architecture. The input passes through a depthwise convolution, after which the resulting feature map is split into sub-feature maps. Convolutions are applied to each sub-feature map, and the outputs are concatenated to form the final feature representation.

The block commences with a depthwise separable convolution. This configuration aims to produce a subnetwork that enhances accuracy while minimizing parameter count and computational complexity. The outcome is subjected to a hard_sigmoid activation, which serves as a non-smooth alternative to the sigmoid function.

Second, to preserve the diversity of feature channels, the incoming feature map X is partitioned into N sub-feature maps. Each sub-feature map comprises several channels, where $N \in \{8, 16, 32\}$ as specified in Section XCC-Net.

Subsequently, the values of the subsequent sub-feature maps are computed using the following formula (Equation (7)), where the input sub-feature map is denoted by SFM and the kernel by K . The indices of the rows and columns of the resulting matrix are represented by l and c respectively. The final outputs Y_i are then passed through a ReLU activation function and concatenated according to Equation (8).

$$Y = (SFM \times K)_{l,c} = \sum_j \sum_k K_{j,k} \times SFM_{l-j,c-k} \quad (7)$$

$$O = \text{Concatenate}(\{\text{ReLU}(Y_i) / i \in \{1, \dots, N\}\}) \quad (8)$$

Nevertheless, the MCSE subnetwork comprises blocks that incorporate bilinear interpolation followed by additional MCSE blocks.

3.4. Powered-Feature Engagement Module

The Powered-Feature Engagement (PFE) module selectively amplifies informative feature responses and suppresses less relevant activations. This enhances feature discriminability before deeper processing. In Figure 4, the PFE module is depicted, comprising three sub-blocks. The first sub-block conducts a depthwise separable convolution. This highlights regions of interest within every channel of the feature map.

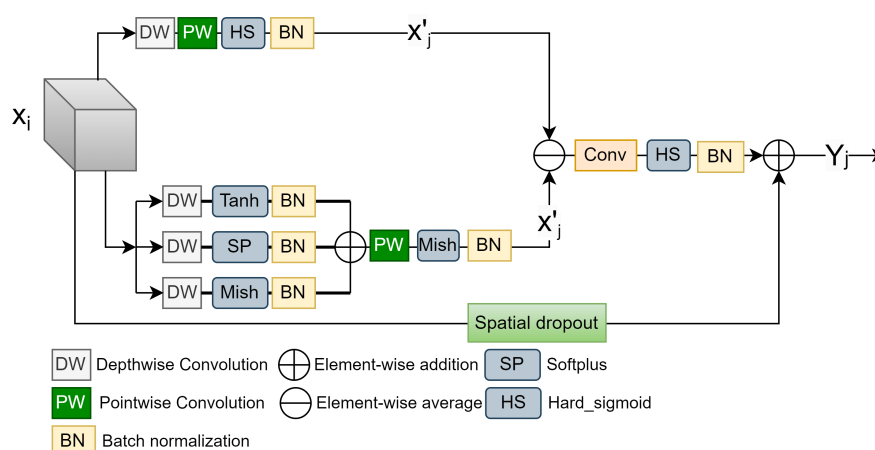


Figure 4. Summary of the proposed PFE module. The first stage has two paths: the first passes through a DW separable convolution, and the second through three DW convolutions followed by a PW convolution. The outputs are averaged, passed through a convolution, and then summed with a skip connection from the input.

A *hard_sigmoid* activation function is then applied to this result. The second sub-block undergoes a triple-fold depthwise separable convolution.

The outputs of both sub-blocks are averaged before being passed through a standard convolution. Furthermore, the third sub-block involves a spatial dropout mechanism. It ensures that the subsequent block in the XSE subnetwork is not overwhelmed by deep features.

Subsequently, the outputs generated by both pathways are averaged and processed through a final convolutional layer. This layer applies a *hard_sigmoid* activation to stabilize the feature representation.

3.5. Global-Feature Ensembling Module

The bottleneck GFE module plays a crucial role as an intermediary component bridging the encoders and decoders. This strategic integration enhances the model's ability to perceive intricate spatial relationships and contextual cues within the input data.

As illustrated in Figure 5, the incoming feature map is processed through three sub-blocks. Element-wise multiplication is employed on the outputs from the three sub-blocks to allow for fine-grained control over the feature representation.

Subsequently, the resulting feature map is further refined through another separable block. Following this, distinct activation functions are applied, allowing for a more comprehensive representation of the underlying features. The outputs from the activation functions are then combined and passed through a Gaussian noise layer. This introduces controlled randomness to the feature map.

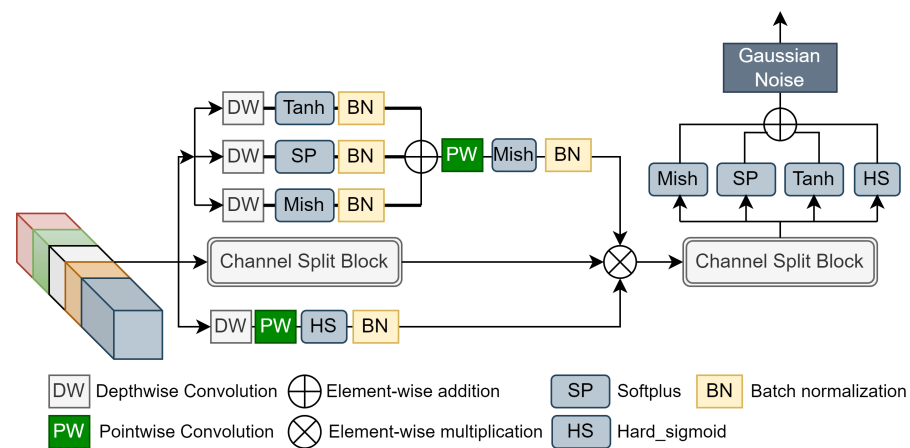


Figure 5. Overview of the proposed bottleneck GFE module. The input passes through three sub-blocks combined via element-wise multiplication. The resulting feature map is forwarded to a channel split block, followed by distinct activation functions. All activation outputs are summed and then passed through a Gaussian noise layer.

4. Experiments

4.1. Datasets

To assess the performance of the XCC-Net architecture, we conducted experiments on multiple datasets, including ISIC 2017 [14], PH2 [15], CVC-ClinicDB [16], and MICCAI 2017 (Red Lesion) [13]. These datasets were chosen to provide a diverse range of lesion types, imaging modalities, and clinical scenarios.

To ensure a consistent evaluation across all datasets, all images were preprocessed using a unified protocol. Specifically, each image was resized to 256×256 pixels and normalized to the range $[0, 1]$. For the CVC-ClinicDB dataset, additional data augmentation was applied to the training set, as described in the CVC-ClinicDB dataset description.

- **ISIC 2017 dataset:** This dataset is a notable contribution, offering a training set comprising 2000 skin lesion images. Alongside these images are corresponding masks for segmentation, superpixel masks for dermoscopic feature extraction, and annotations for classification purposes. This dataset encompasses lesions categorized into melanoma, seborrheic keratosis, and nevus, totaling 2750 images. Among these, 2000 images are included in the training subset, 150 images in validation, and 600 images in testing.
- **PH2 dataset:** The PH2 dataset [15] includes 200 images, predominantly featuring naevus (160 images) and melanoma (40 images). These images are all 8-bit RGB with a resolution of 768×560 pixels, captured using a $20\times$ magnification lens. The PH2 dataset is exclusively reserved for testing purposes.
- **MICCAI 2017 dataset:** The MICCAI 2017 [13] dataset contains 3895 frames, each sized at 320×320 pixels. It encompasses both normal frames and those featuring identified lesions such as bleeding and angioectasias. The dataset was partitioned, allocating 80% of the data for training and reserving 20% for testing and validation.
- **CVC-ClinicDB:** The CVC-ClinicDB dataset [16] comprises 612 images sourced from 31 colonoscopy sequences. Each image is sized at 384×288 pixels. To facilitate model training and evaluation, the dataset is divided into training and testing sets. Specifically, 82% of the data is allocated for training, while 18% is reserved for testing purposes. For enhanced training robustness, the training set is augmented through rotations, flips, and brightness adjustments, resulting in a total of 2000 images.

4.2. Experiment Setup

4.2.1. Implementation Details

The XCC-Net model is implemented using the TensorFlow framework. The computer is equipped with an NVIDIA RTX 3090 24 GB GDDR6X graphics card. Prior to training, the images are resized to a resolution of 256×256 pixels to suit the model architecture. To optimize the network, we employ the Adamax optimizer [49].

We set the final learning rate to 0.001 and chose values of 0.9 and 0.999 for β_1 and β_2 , respectively. During training, we utilized a Dice loss function along with a batch size of 5. Training for just 150 epochs produced satisfactory results. The Dice loss is computed as follows:

$$L_{dice} = \frac{2 \times \sum Y_{true} \times Y_{pred}}{\sum Y_{true}^2 + \sum Y_{pred}^2} \quad (9)$$

where Y_{true} represents the pixel values in the ground truth segmentation map and Y_{pred} denotes the predicted pixel values.

Regarding regularization, we applied Gaussian noise within the GFE module to enhance robustness against feature perturbations, while spatial dropout was employed in the PFE module to reduce feature co-adaptation and improve generalization. We did not use early stopping or learning-rate scheduling during the main training procedure. However, initial learning rate values were explored within the ablation study to determine the optimal hyperparameter configuration for XCC-Net.

All standard baseline models were reimplemented and trained under a unified pipeline. For recent state-of-the-art models, we reported the results as published in their original works, without modifying their configurations. This approach allows for meaningful performance evaluation by avoiding biases that could result from inconsistent training settings or architectural modifications.

4.2.2. Evaluation Metrics

Intersection over Union (IoU) and Dice coefficient (DC), both widely utilized metrics for evaluating performance, are employed in this study. The IoU is computed by dividing the area of overlap between the predicted segmentation map (B) and the ground truth (A) by the total area encompassed by the two regions. The DC is calculated by multiplying the area of overlap by two and then dividing it by the sum of the pixels in both images. The formulas for computing IoU and DC are as follows:

$$IoU = \frac{|A \cap B|}{|A \cup B|} \quad (10)$$

$$DC = \frac{2 \times |A \cap B|}{|A| + |B|} \quad (11)$$

Floating Point Operations (FLOPs) is a standard computational complexity metric that quantifies the total number of floating-point arithmetic operations a neural network requires to process a single input. When expressed in billions of operations, it is referred to as Giga FLOPs (GFLOPs). FLOPs provide a hardware-independent measure of model efficiency and helps compare the computational cost between different architectures.

For CNNs, the FLOPs for a single convolutional layer can be estimated as:

$$\text{FLOPs}_{\text{conv}} = 2 \times (K_h \times K_w \times C_{\text{in}}) \times (H_{\text{out}} \times W_{\text{out}} \times C_{\text{out}}) \quad (12)$$

where K_h and K_w denote the convolution kernel height and width, C_{in} and C_{out} represent the number of input and output channels, while H_{out} and W_{out} correspond to the spatial dimensions of the resulting output feature map. The factor 2 is included to account for

both multiplication and addition operations performed in each Multiply–Accumulate (MAC) computation.

The total FLOPs of the network is obtained by summing the FLOPs of every computational layer:

$$\text{FLOPs}_{\text{total}} = \sum_{l=1}^L \text{FLOPs}_l \quad (13)$$

4.3. Results

We evaluated the proposed method against several state-of-the-art models on the publicly available datasets mentioned. All baseline models, including U-Net, U-Net++, and SegNet, were trained from scratch on all datasets using their original architectures. No external pretraining or architectural modifications were applied. The dataset poses inherent challenges due to high lesion variability, complex backgrounds, and diverse imaging conditions.

4.3.1. Results on MICCAI 2017 Dataset (Red Lesion)

Table 1 presents the performance metrics on the MICCAI 2017 dataset, specifically the DC and IoU, of various segmentation models including SegNet [25], V-Net [50], ResUNet-a [20], ResUNet++ [51], TernaUSNet [19], Attention UNet [18], UNet++ [17], UNet [4], and the proposed method. All the networks evaluated in this comparison were subjected to identical conditions to ensure a fair and reliable assessment.

The findings presented in Table 1 illustrate that our model surpasses the currently available approaches. In comparison to other segmentation techniques, our model achieved strong results, boasting a DC of 91.7% and an IoU score of 84.68%.

Table 1. Comparisons of the proposed segmentation technique with advanced segmentation methods using the MICCAI 2017 dataset. Bold text indicates the best-performing result in each row or column, while a dash (–) denotes unavailable data.

Method	DC (%)	IoU (%)	Param (M)	GFLOPs
SegNet [25]	72.37	56.71	11.74	51.04
U-Net [4]	72.37	56.71	34.51	101.75
U-Net++ [17]	80.45	67.29	9.04	59.6
Charfi et al. [52]	80.66	71.29	–	–
Attention U-Net [18]	90.16	82.09	35.23	135.14
V-Net [50]	90.54	82.72	23.75	21.76
TernaUSNet [19]	90.67	82.93	23.01	62.33
TransCS-Net [53]	91.11	85.52	51.97	58.33
Garbaz et al. [54]	91.14	83.72	14.92	27.47
ResUNet++ [51]	91.18	83.80	4.07	32.07
ResUNet-a [20]	91.53	84.38	6.28	23.99
XCC-Net	91.7	84.68	22.55	6.05

The encoder utilized in the approach presented by [19] was based on a pre-trained VGG11 architecture, excluding its fully connected layers. We opted to apply their system without the pre-trained weights to ensure a more equitable comparison.

Moreover, the performance of the attention U-Net [18] was notably favorable, owing to its utilization of spatial regions identified through the examination of activations and context-related data. This approach achieved an IoU of 82.93% and a DC of 90.67%.

Similarly, U-Net++ [17] enhanced skip connections by amalgamating features with diverse spatial hierarchies in the decoder block, resulting in a DC of 80.45%. On the other

hand, the straightforward architectures of U-Net [4] and SegNet [25] yielded comparable outcomes.

Indeed, V-Net [50], ResUNet-a [20], and ResUNet++ [51] exhibit remarkable performance, boasting DCs surpassing 0.90 and IoU values exceeding 0.82. This indicates their efficacy in accurately segmenting lesions from medical images. These results highlight the competitiveness of these models and their potential for reliable lesion segmentation tasks in medical imaging.

Finally, XCC-Net demonstrates a computational complexity of only 6.05 GFLOPs, which is significantly lower compared to other methods. This lower computational requirement indicates that XCC-Net is more efficient and faster, making it suitable for real-time or resource-constrained applications, without sacrificing performance.

4.3.2. Results on ISIC 2017 and PH2 Dataset

Table 2 presents a comparative analysis of XCC-Net's performance against other advanced segmentation techniques using the ISIC 2017 dataset [14]. Overall, U-Net [4], Attention U-Net [18], the proposed method, V-Net [50], and ResUNet-a [20] stand out as the top-performing models. This indicates their suitability for segmentation tasks requiring high accuracy and strong overlap with ground truth annotations.

Table 2. The performance of various methods on the ISIC 2017 and PH2 datasets.

Method	ISIC 2017		PH2	
	DC (%)	IoU (%)	DC (%)	IoU (%)
U-Net++ [17]	29.50	18.06	37.29	23.44
TernausNet [19]	38.19	23.60	48.78	32.25
SegNet [25]	48.78	32.25	48.78	32.25
U-Net [4]	77.70	63.54	87.14	78.21
ResUNet-a [20]	78.63	64.78	88.47	80.12
Attention U-Net [18]	79.01	65.30	85.83	75.89
V-Net [50]	79.75	66.33	88.77	80.64
ResUNet++ [51]	82.26	69.87	87.63	78.35
PCCTrans [55]	84.65	-	-	-
FAT-Net [56]	85	76.35	-	-
XCC-Net	79.07	65.39	89.26	81.30

XCC-Net exhibits a more complex architecture than U-Net and Attention U-Net, incorporating dual encoders and decoders, which contributes to its competitive performance. The superior performance of Attention U-Net [18] can be attributed to its attention mechanisms incorporated into the skip connections, actively suppressing activations in irrelevant regions.

In comparison, XCC-Net surpasses attention U-Net and U-Net by 0.6% and 1.37%, respectively, in terms of DC. However, the lower metric scores of the other methods may be partially attributed to architectural limitations that affect generalization across diverse segmentation scenarios.

V-Net [50] achieved higher performance than XCC-Net due to its capability to directly process volumetric data and capture spatial dependencies across multiple dimensions. Similarly, ResUNet++ [51] achieved better results due to its adeptness in effectively capturing contextual information through a multi-scale feature fusion mechanism. Despite these strong results, XCC-Net demonstrates better generalization performance when evaluated on unseen datasets. However, in the evaluation conducted on the unseen PH2 dataset, XCC-Net surpassed V-Net and ResUNet++ by 0.49% and 1.63% in DC respectively, indicating the superior generalizability of the proposed XCC-Net architecture.

However, XCC-Net shows a performance drop of 5.93% compared to the best model FAT-Net [56]. Despite this lower accuracy, XCC-Net requires only 6.05 GFLOPs, which is significantly less than the 23 GFLOPs required by FAT-Net. Furthermore, XCC-Net shows a substantial reduction in both computational complexity and the number of parameters compared to PCCTrans [55], which has 50.8 million parameters and 38.5 GFLOPs.

To further assess the robustness and cross-dataset performance of our method, we conducted evaluations on the PH2 dataset. The model, originally trained on the ISIC 2017 dataset, was applied to the PH2 dataset, which served as unseen data during training. As depicted in Table 2, the results demonstrate that XCC-Net achieved the highest DC of 89.26% and IoU of 81.3%.

4.3.3. Results on CVC-ClinicDB Dataset

Similarly, XCC-Net was evaluated on the CVC-ClinicDB dataset, as detailed in Table 3. This dataset was used to assess the performance of the proposed method in segmenting polyps in colonoscopy images. The results indicate that XCC-Net achieved a DC of 87.15% and IoU of 77.23%, demonstrating its effectiveness in polyp segmentation tasks within colonoscopy images. On this dataset, ResUNet++ [51] reported a mean IoU of 79.62%. In contrast, the IoU metric in this study is computed globally. Nevertheless, it remains a directly comparable indicator of performance in the binary segmentation setting.

Additionally, XCC-Net outperforms transformer-based methods MBP-SSNet [57] and TranSEFusionNet [58] by margins of 0.58% and 0.67% in DC, respectively. Moreover, the proposed method exhibits a substantial reduction in computational complexity, requiring only 6.05 GFLOPs compared to 105.6 GFLOPs for MBP-SSNet [57].

Table 3. Comparisons of the effectiveness of different techniques in polyp detection.

Method	DC (%)	IoU (%)	Param (M)	GFLOPs
ResUNet++ [51]	79.55	79.62	4.07	32.07
V-Net [50]	79.59	66.10	23.75	21.76
U-Net [4]	81.10	68.21	34.51	101.75
ResUNet-a [20]	84.32	72.89	6.28	30.04
Attention U-Net [18]	84.38	72.99	35.23	135.14
TranSEFusionNet [58]	86.48	79.09	127.74	124.43
MBP-SSNet [57]	86.57	78.24	9.37	105.6
XCC-Net	87.15	77.23	22.55	6.05

4.4. Ablation Study

4.4.1. Ablation Study of Subnetworks

We conducted comprehensive ablation studies (Table 4) to evaluate the effectiveness of the subnetworks and modules within XCC-Net. We focused in particular on two baseline subnetworks: XSE+XSD and MCSE+MCSD. The XSE+XSD subnetworks exhibited significant performance metrics, achieving a DC of 90.89% and an IoU of 83.83.

These impressive results can be attributed to the inherent capability of XSE blocks to extract precise local and positional information. Additionally, the XSE blocks minimize the loss of spatial context. On the other hand, the MCSE+MCSD configuration achieved respectable performance, with a DC of 88.39 and an IoU of 79.83%.

These metrics, while slightly lower than those of XSE+XSD, are still commendable. They outperform traditional methods that rely solely on skip connections between encoding and decoding paths.

The MCSE component was separately evaluated. This evaluation aimed to assess its effectiveness in progressively fusing semantic and structural information from deep

features. Furthermore, we attempted to combine the advantages of both XSE+XSD and MCSE+MCSD subnetworks by merging them into an X-shaped architecture.

However, this combined approach did not outperform the XSE+XSD network (Figure 6). The discrepancy in performance can be attributed to the divergent information from both subnetworks. These subnetworks converge at the bottleneck GFE before being processed again by XSD and MCSD. This convergence leads to the generation of ambiguous or noisy feature maps.

Table 4. Analysis of the impact of subnetworks and modules through an ablation study conducted on the MICCAI 2017 dataset. ✓ indicates the presence of the corresponding feature.

No	MCSE	GFE	XSE	PFE	Encoder–Decoder Skip Connections	DC (%)	IoU (%)
A	✓	✓				88.39	79.83
B	✓	✓	✓			90.03	82.42
C	✓	✓	✓	✓	✓	90.2	82.15
D		✓	✓			90.89	83.83
E	✓	✓	✓	✓		91.7	84.68

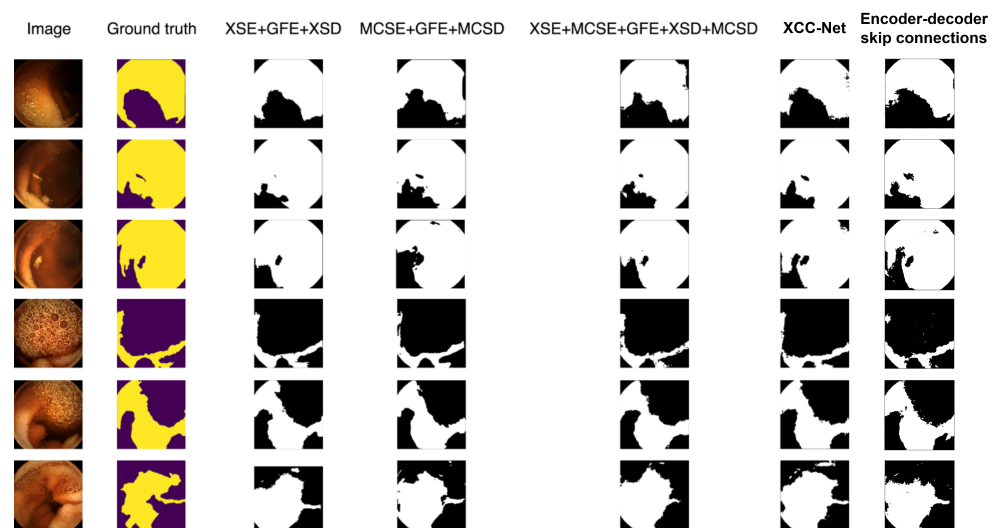


Figure 6. Visual representation showcasing the segmentation outcomes across various subnetworks and modules on the MICCAI 2017 dataset. The XSD+GFE+XSD configuration achieves a DC of 90.89% and IoU of 83.83%. The MCSE+GFE+MCSD configuration yields a DC of 88.39% and IoU of 79.83%. The combination XSE+MCSE+GFE+XSD+MCSD reaches a DC of 90.03% and IoU of 82.42%. When all components are combined, the model achieves the highest performance with a DC of 91.70% and IoU of 84.68%. The variant of the proposed model that includes encoder–decoder skip connections achieved a DC of 90.20% and an IoU of 82.15%. White and black areas indicate the bleeding region and surrounding tissue, respectively.

To address this limitation, we introduced the PFE module between each XSE and MCSE encoder. This module facilitates the transfer of information from MCSE to XSE blocks. It enables the forwarding of unseen regions of interest. Additionally, it stabilizes feature maps from both encoder subnetworks.

Moreover, we standardized the number of channels to 64, 128, 256, and 512 for all subnetworks, with 1024 channels for the GFE.

Finally, to demonstrate the effectiveness of the proposed method, we included an additional ablation experiment on encoder–decoder skip connections. In this variant, we integrated skip connections between the XSE encoder and its corresponding XSD decoder, where the most informative features propagate.

As shown in Table 4, the model incorporating these encoder–decoder skip connections achieves a DC of 90.20% and an IoU of 82.15%. This performance is lower than that of the full XCC-Net. The decrease can be explained by the need for spatial cropping to compensate for border pixel loss caused by convolutions. This cropping reduces valuable contextual information and negatively affects segmentation accuracy. The same behavior can be observed in Figure 6, specifically in the column corresponding to the encoder–decoder skip connections variant.

Figure 6 illustrates the lesion segmentation results for various subnetwork and module combinations on the red lesion dataset. It is evident from the results that each component of XCC-Net contributes to progressively refining the segmentation map. This demonstrates the effectiveness of the integrated architecture.

4.4.2. Ablation Study of Hyperparameters

To analyze the influence of hyperparameters on segmentation performance, we conducted an ablation study on the MICCAI 2017 dataset. We varied the learning rate (α) and the Adamax optimizer coefficients β_1 and β_2 . The baseline configuration (Model A) was selected as a conservative starting point with relatively low hyperparameter values. The quantitative results of these experiments are summarized in Table 5.

Among all evaluated configurations, the setting $\alpha = 0.001$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$ (Model D) achieved the highest performance, yielding a DC of 91.70% and an IoU of 84.68%. It consistently outperformed the other parameter combinations, demonstrating superior convergence behavior and segmentation accuracy.

Lower hyperparameter values, as in Model A, resulted in reduced segmentation performance. Increasing the learning rate and optimizer coefficients, as in Model B, provided moderate improvements but remained inferior to Model D. The configuration in Model C led to a noticeable degradation in both Dice and IoU scores. Based on these observations, we selected the hyperparameter setting of Model D for all subsequent experiments.

Figure 7 presents qualitative segmentation results under different hyperparameter combinations. Models A, B, and C occasionally fail to capture small normal regions within bleeding zones. This indicates limited sensitivity to fine-grained boundaries.

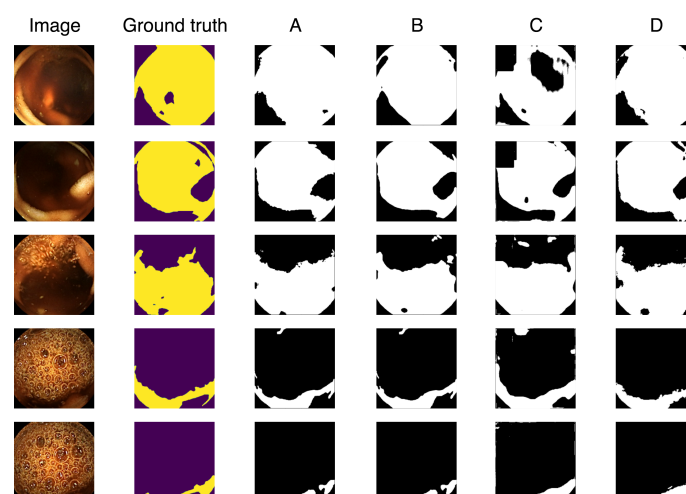


Figure 7. Visualization illustrating segmentation results under different hyperparameter configurations on the MICCAI 2017 dataset. Model A achieves a DC of 89.35% and an IoU of 81.48%. Model B achieves 90.22% DC and 82.73% IoU, while Model C records 87.68% DC and 78.06% IoU. Model D shows the best performance with 91.70% DC and 84.68% IoU. The white area highlights the bleeding region. The black area represents the surrounding tissue.

In contrast, Model D produces segmentation maps that more closely resemble the ground truth annotations. Minor misclassifications still persist in visually ambiguous regions where normal tissue shares similar appearances with bleeding areas. This highlights the inherent challenge of precise lesion boundary segmentation.

Table 5. Analysis of different combinations of hyperparameters through an ablation study conducted on the MICCAI 2017 dataset.

NO	α	β_1	β_2	DC (%)	IoU (%)
A	0.0001	0.8	0.89	89.35	81.48
B	0.01	0.99	1	90.22	82.73
C	0.01	0.9	0.89	87.68	78.06
D	0.001	0.9	0.999	91.70	84.68

4.5. Discussion

U-Net variants often incorporate skip connections, a key feature inherited from the original U-Net architecture, to preserve spatial information during encoding and aid in the recovery of finer details during decoding.

While skip connections are beneficial for segmentation accuracy, they can obscure the model's decision-making process by establishing direct connections between encoder and decoder layers. This opacity makes it difficult to interpret why certain regions are segmented in specific ways.

It can hinder the model's generalizability to diverse tasks and lesion types. To overcome these limitations, we introduced XCC-Net. It is a lightweight model that combines various subnetworks. This combination enhances segmentation performance while maintaining generalizability.

We evaluated XCC-Net across multiple medical image segmentation datasets. As illustrated in Figures 8 and 9, XCC-Net demonstrates robust segmentation of diverse lesions across different datasets. It performs well even when dealing with unseen data, such as in the case of ISIC 2017 and PH2.

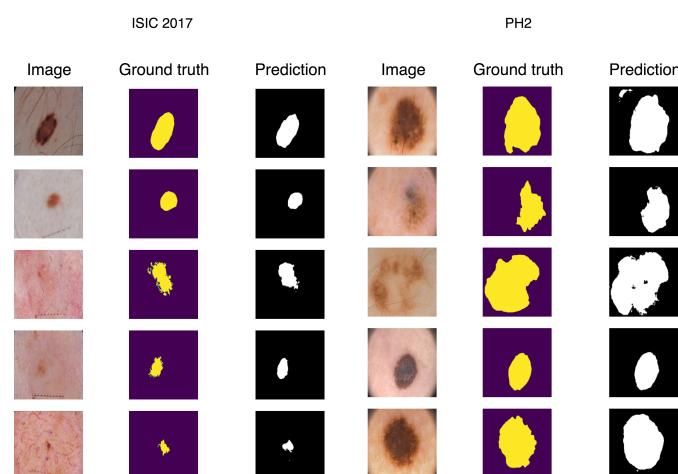


Figure 8. Visualization of XCC-Net segmentation results on the ISIC 2017 and PH2 datasets. On ISIC 2017, the model achieved a DC of 79.07% and an IoU of 65.39%. On the PH2 dataset, it achieved a DC of 89.26% and an IoU of 81.30%. White and black areas represent the skin lesion region and the surrounding tissue, respectively.

However, our approach faces two primary challenges. First, while the incorporation of element-wise operations and diverse subnetwork combinations improves performance, it introduces computational inefficiencies that can affect inference speed, particularly in real-

time applications. Second, accurately segmenting low-contrast lesions and distinguishing them from surrounding tissue remains difficult.

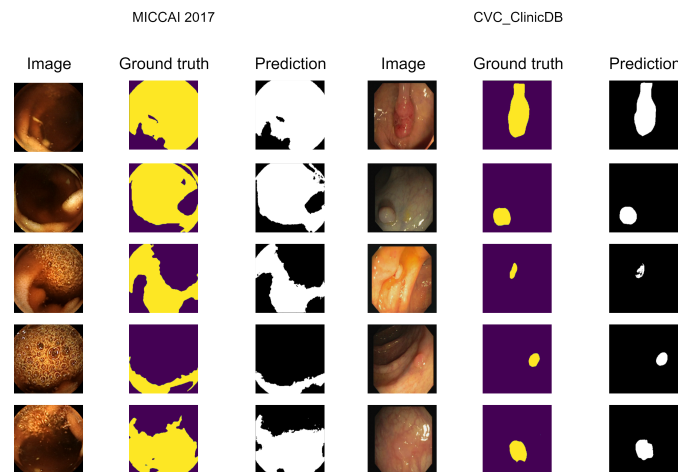


Figure 9. Visualization of XCC-Net segmentation performance on the MICCAI 2017 (red lesion) and CVC-ClinicDB datasets. On MICCAI 2017, the model achieved a DC of 91.70% and an IoU of 84.68%. On the CVC-ClinicDB dataset, it achieved a DC of 87.15% and an IoU of 77.23%. The white area indicates the polyp and bleeding regions. The black area represents the surrounding tissue.

This is evident in predictions from ISIC 2017 and MICCAI 2017 datasets in Figures 8 and 9, respectively.

This issue arises from the lack of a consistent structural relationship between the two decoders. Combining their outputs provides richer and more complementary feature representations, since each decoder can capture different semantic cues.

However, in the subsequent processing steps, one decoder may dominate the other, causing some informative features to be overlooked or underutilized. This issue arises from the lack of a consistent structural relationship between the two decoders.

This behavior differs from the typical skip connections employed in conventional encoder–decoder architectures. As a result, the model may not fully exploit the interactions between the two decoders, which can hinder segmentation performance, especially in regions with subtle boundaries or low contrast. Low-contrast lesions create indistinct regions, making it difficult for the model to accurately detect and separate them from surrounding tissue.

In future work, we aim to refine the integration of the subnetworks by optimizing element-wise merging operations and simplifying the combination of modules. We also plan to introduce a more efficient fusion mechanism between the decoder branches. These improvements are expected to enhance the segmentation of low-contrast lesions and improve the model’s ability to differentiate lesion regions from surrounding tissues.

5. Conclusions

This paper introduces a novel method, called XCC-Net, for medical image segmentation. XCC-Net employs skip connections only between the two encoders. This design choice omits direct connections between encoders and decoders. It also aims to improve interpretability. Direct connections can obscure the rationale behind the model’s segmentation decisions, hindering interpretability.

XCC-Net is engineered to detect various types of lesions across multiple imaging modalities, including WCE, colonoscopy, and dermoscopic images. It consists of two integrated subnetwork branches: XSE-XSD and MCSE-MCSD. Additionally, it incorporates a Bottleneck GFE module and a PFE module integrated between each pair of en-

coder blocks. Extensive experiments demonstrate the efficacy of XCC-Net across diverse segmentation tasks.

Despite its effectiveness, XCC-Net faces two primary challenges. Firstly, the combination of different subnetworks enhances performance but incurs computational overheads, affecting inference speed. Secondly, the model struggles with segmenting low-contrast lesions. It also faces challenges in distinguishing between lesions and their surroundings. This struggle arises due to the lack of connections between decoders to ensemble and stabilize feature maps.

To address these limitations, future work will focus on refining the combination of subnetworks. This will involve optimizing and minimizing module combinations. Additionally, an additional fusion module between decoders will be introduced. This module aims to improve segmentation accuracy for low-contrast lesions. It also seeks to enhance differentiation between lesions and their surroundings.

Author Contributions: Conceptualization, A.G., Y.O., S.C., M.E.A. and L.K.; Methodology, A.G. and Y.O.; Software, L.K. and M.E.A.; Validation, A.G., Y.O., S.C., M.E.A., L.K., A.M.L., M.H. and M.O.; Formal analysis, A.G., Y.O., S.C., M.E.A., L.K., A.M.L., M.H. and M.O.; Investigation, A.G., Y.O., S.C., M.E.A., L.K., A.M.L., M.H. and M.O.; Resources, L.K., M.E.A. and M.H.; Data curation, S.C. and M.E.A.; Writing—original draft preparation, A.G., Y.O., S.C., M.E.A. and L.K.; Writing—review and editing, A.G., Y.O., S.C., M.E.A., L.K., A.M.L., M.H. and M.O.; Visualization, A.G. and Y.O.; Supervision, M.E.A. and L.K.; Project administration, L.K.; Funding acquisition, L.K., M.E.A. and M.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Ministry of National Education by Vocational Training; in part by the Higher Education and Scientific Research through the Ministry of Industry, Trade, and Green and Digital Economy; in part by the Digital Development Agency (ADD); and in part by the National Center for Scientific and Technical Research (CNRST) under Project ALKHAWARIZMI/2020/20.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets analysed during the current study are available in the Red Lesion Endoscopy repository (<https://rdm.inesctec.pt/dataset/nis-2018-003>, accessed on 6 December 2025), the ISIC Challenge Datasets repository (<https://challenge.isic-archive.com/data/>, accessed on 6 December 2025), and the CVC-ClinicDB repository (<https://www.kaggle.com/datasets/balraj98/cvcclinicdb>, accessed on 6 December 2025).

Acknowledgments: The authors have reviewed and edited the output and take full responsibility for the content of this publication. The implementation of XCC-Net used in this study is available from the authors upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional Neural Network
GFE	Global-Feature Ensembling bottleneck module
MCSD	Multi-Channel Separable Decoder
MCSE	Multi-Channel Separable Encoder
PFE	Powered-Feature Engagement module
WCE	Wireless Capsule Endoscopy
XSD	X-Separable Decoder
XSE	X-Separable Encoder

References

1. Minaee, S.; Boykov, Y.; Porikli, F.; Plaza, A.; Kehtarnavaz, N.; Terzopoulos, D. Image segmentation using deep learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3523–3542. [\[CrossRef\]](#)
2. Hesamian, M.H.; Jia, W.; He, X.; Kennedy, P. Deep learning techniques for medical image segmentation: Achievements and challenges. *J. Digit. Imaging* **2019**, *32*, 582–596. [\[CrossRef\]](#)
3. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
4. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
5. Siegel, R.L.; Miller, K.D.; Wagle, N.S.; Jemal, A. Cancer statistics, 2023. *CA Cancer J. Clin.* **2023**, *73*, 17–48. [\[CrossRef\]](#)
6. Linares, M.A.; Zakaria, A.; Nizran, P. Skin cancer. *Prim. Care Clin. Off. Pract.* **2015**, *42*, 645–659. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Schadendorf, D.; Fisher, D.E.; Garbe, C.; Gershenwald, J.E.; Grob, J.J.; Halpern, A.; Herlyn, M.; Marchetti, M.A.; McArthur, G.; Ribas, A.; et al. Melanoma. *Nat. Rev. Dis. Prim.* **2015**, *1*, 15003. [\[CrossRef\]](#)
8. Ring, C.; Cox, N.; Lee, J.B. Dermatoscopy. *Clin. Dermatol.* **2021**, *39*, 635–642. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Xie, Y.; Shi, L.; He, X.; Luo, Y. Gastrointestinal cancers in China, the USA, and Europe. *Gastroenterol. Rep.* **2021**, *9*, 91–104. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Kim, B.S.M.; Li, B.T.; Engel, A.; Samra, J.S.; Clarke, S.; Norton, I.D.; Li, A.E. Diagnosis of gastrointestinal bleeding: A practical guide for clinicians. *World J. Gastrointest. Pathophysiol.* **2014**, *5*, 467–478. [\[CrossRef\]](#)
11. Fisher, D.A.; Maple, J.T.; Ben-Menachem, T.; Cash, B.D.; Decker, G.A.; Early, D.S.; Evans, J.A.; Fanelli, R.D.; Fukami, N.; Hwang, J.H.; et al. Complications of colonoscopy. *Gastrointest. Endosc.* **2011**, *74*, 745–752. [\[CrossRef\]](#)
12. Iddan, G.; Meron, G.; Glukhovskiy, A.; Swain, P. Wireless capsule endoscopy. *Nature* **2000**, *405*, 417. [\[CrossRef\]](#)
13. Coelho, P.; Pereira, A.; Leite, A.; Salgado, M.; Cunha, A. A deep learning approach for red lesions detection in video capsule endoscopies. In Proceedings of the Image Analysis and Recognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, 27–29 June 2018; Proceedings 15; Springer: Berlin/Heidelberg, Germany, 2018; pp. 553–561.
14. Codella, N.C.; Gutman, D.; Celebi, M.E.; Helba, B.; Marchetti, M.A.; Dusza, S.W.; Kalloo, A.; Liopyris, K.; Mishra, N.; Kittler, H.; et al. 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 Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 4–7 April 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 168–172.
15. Mendonça, T.; Ferreira, P.M.; Marques, J.S.; Marcal, A.R.; Rozeira, J. PH 2-A dermoscopic image database for research and benchmarking. In Proceedings of the 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka, Japan, 3–7 July 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 5437–5440.
16. Bernal, J.; Sánchez, F.J.; Fernández-Esparrach, G.; Gil, D.; Rodríguez, C.; Vilariño, F. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Comput. Med. Imaging Graph.* **2015**, *43*, 99–111. [\[CrossRef\]](#)
17. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging* **2019**, *39*, 1856–1867. [\[CrossRef\]](#)
18. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention u-net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999. [\[CrossRef\]](#)
19. Iglovikov, V.; Shvets, A. Ternaunet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. *arXiv* **2018**, arXiv:1801.05746.
20. Diakogiannis, F.I.; Waldner, F.; Caccetta, P.; Wu, C. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 94–114. [\[CrossRef\]](#)
21. Lafraxo, S.; Souaidi, M.; El Ansari, M.; Koutti, L. Semantic segmentation of digestive abnormalities from wce images by using attresu-net architecture. *Life* **2023**, *13*, 719. [\[CrossRef\]](#)
22. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306. [\[CrossRef\]](#)
23. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 205–218.
24. Alom, M.Z.; Hasan, M.; Yakopcic, C.; Taha, T.M.; Asari, V.K. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. *arXiv* **2018**, arXiv:1802.06955.
25. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [\[CrossRef\]](#)

26. Jia, X.; Xing, X.; Yuan, Y.; Xing, L.; Meng, M.Q.H. Wireless capsule endoscopy: A new tool for cancer screening in the colon with deep-learning-based polyp recognition. *Proc. IEEE* **2019**, *108*, 178–197. [\[CrossRef\]](#)
27. Borgli, H.; Stensland, H.K.; Halvorsen, P. Automatic prompt generation using class activation maps for foundational models: A polyp segmentation case study. *Mach. Learn. Knowl. Extr.* **2025**, *7*, 22. [\[CrossRef\]](#)
28. Charfi, S.; EL Ansari, M.; Koutti, L.; ELjaafari, I.; ELLahyani, A. Feature Pyramid Network Based Spatial Attention and Cross-Level Semantic Similarity for Diseases Segmentation From Capsule Endoscopy Images. *Int. J. Imaging Syst. Technol.* **2024**, *34*, e23194. [\[CrossRef\]](#)
29. Souaidi, M.; Lafraxo, S.; Kerkaou, Z.; El Ansari, M.; Koutti, L. A multiscale polyp detection approach for gi tract images based on improved densenet and single-shot multibox detector. *Diagnostics* **2023**, *13*, 733. [\[CrossRef\]](#)
30. Ellahyani, A.; Jaafari, I.E.; Charfi, S.; Ansari, M.E. Fine-tuned deep neural networks for polyp detection in colonoscopy images. *Pers. Ubiquitous Comput.* **2023**, *27*, 235–247. [\[CrossRef\]](#)
31. Lafraxo, S.; El Ansari, M.; Koutti, L. Gastrosegnet: Polyp segmentation using colonoscopic images based on attentionu-net architecture. In Proceedings of the 2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM), Istanbul, Turkey, 26–28 October 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–6.
32. Ghosh, T.; Fattah, S.A.; Wahid, K.A. CHOBS: Color histogram of block statistics for automatic bleeding detection in wireless capsule endoscopy video. *IEEE J. Transl. Eng. Health Med.* **2018**, *6*, 1800112. [\[CrossRef\]](#)
33. Caroppo, A.; Leone, A.; Siciliano, P. Deep transfer learning approaches for bleeding detection in endoscopy images. *Comput. Med Imaging Graph.* **2021**, *88*, 101852. [\[CrossRef\]](#)
34. Kanakatte, A.; Ghose, A. Precise Bleeding and Red lesions localization from Capsule Endoscopy using Compact U-Net. In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Virtual, 1–5 November 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 3089–3092.
35. Bai, F.; Xing, X.; Shen, Y.; Ma, H.; Meng, M.Q.H. Discrepancy-based active learning for weakly supervised bleeding segmentation in wireless capsule endoscopy images. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Singapore, 18–22 September 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 24–34.
36. Li, S.; Zhang, J.; Ruan, C.; Zhang, Y. Multi-stage attention-unet for wireless capsule endoscopy image bleeding area segmentation. In Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, 18–21 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 818–825.
37. Hajabdollahi, M.; Esfandiarpour, R.; Najarian, K.; Karimi, N.; Samavi, S.; Soroushmehr, S.R. Low complexity crn structure for automatic bleeding zone detection in wireless capsule endoscopy imaging. In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 7227–7230.
38. Jain, S.; Seal, A.; Ojha, A.; Yazidi, A.; Bures, J.; Tacheci, I.; Krejcar, O. A deep CNN model for anomaly detection and localization in wireless capsule endoscopy images. *Comput. Biol. Med.* **2021**, *137*, 104789. [\[CrossRef\]](#)
39. Xing, X.; Yuan, Y.; Meng, M.Q.H. Zoom in lesions for better diagnosis: Attention guided deformation network for wce image classification. *IEEE Trans. Med. Imaging* **2020**, *39*, 4047–4059. [\[CrossRef\]](#) [\[PubMed\]](#)
40. Mirikharaji, Z.; Abhishek, K.; Bissoto, A.; Barata, C.; Avila, S.; Valle, E.; Celebi, M.E.; Hamarneh, G. A survey on deep learning for skin lesion segmentation. *Med. Image Anal.* **2023**, *88*, 102863.
41. Öztürk, Ş.; Özkaya, U. Skin lesion segmentation with improved convolutional neural network. *J. Digit. Imaging* **2020**, *33*, 958–970. [\[CrossRef\]](#)
42. Xie, F.; Yang, J.; Liu, J.; Jiang, Z.; Zheng, Y.; Wang, Y. Skin lesion segmentation using high-resolution convolutional neural network. *Comput. Methods Programs Biomed.* **2020**, *186*, 105241. [\[CrossRef\]](#)
43. Li, H.; He, X.; Zhou, F.; Yu, Z.; Ni, D.; Chen, S.; Wang, T.; Lei, B. Dense deconvolutional network for skin lesion segmentation. *IEEE J. Biomed. Health Inform.* **2018**, *23*, 527–537. [\[CrossRef\]](#)
44. Lei, B.; Xia, Z.; Jiang, F.; Jiang, X.; Ge, Z.; Xu, Y.; Qin, J.; Chen, S.; Wang, T.; Wang, S. Skin lesion segmentation via generative adversarial networks with dual discriminators. *Med. Image Anal.* **2020**, *64*, 101716. [\[CrossRef\]](#)
45. Wu, H.; Pan, J.; Li, Z.; Wen, Z.; Qin, J. Automated skin lesion segmentation via an adaptive dual attention module. *IEEE Trans. Med. Imaging* **2020**, *40*, 357–370. [\[CrossRef\]](#) [\[PubMed\]](#)
46. Hasan, M.K.; Dahal, L.; Samarakoon, P.N.; Tushar, F.I.; Martí, R. DSNet: Automatic dermoscopic skin lesion segmentation. *Comput. Biol. Med.* **2020**, *120*, 103738. [\[CrossRef\]](#) [\[PubMed\]](#)
47. Zhu, W.; Tian, J.; Chen, M.; Chen, L.; Chen, J. MSS-UNet: A Multi-Spatial-Shift MLP-based UNet for skin lesion segmentation. *Comput. Biol. Med.* **2024**, *168*, 107719. [\[CrossRef\]](#) [\[PubMed\]](#)
48. Sifre, L. Rigid-Motion Scattering for Image Classification. Ph.D. Thesis, 2014.
49. Kinga, D.; Adam, J.B. A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980. [\[CrossRef\]](#)

50. Milletari, F.; Navab, N.; Ahmadi, S.A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 565–571.
51. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; Johansen, D.; De Lange, T.; Halvorsen, P.; Johansen, H.D. Resunet++: An advanced architecture for medical image segmentation. In Proceedings of the 2019 IEEE International Symposium on Multimedia (ISM), San Diego, CA, USA, 9–11 December 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 225–2255.
52. Charfi, S.; Ansari, M.E.; Koutti, L.; Ellahyani, A.; Eljaafari, I. Modified residual attention network for abnormalities segmentation and detection in WCE images. *Soft Comput.* **2024**, *28*, 6923. [[CrossRef](#)]
53. Tang, S.; Cheang, C.F.; Yu, X.; Liang, Y.; Feng, Q.; Chen, Z. TransCS-Net: A hybrid transformer-based privacy-protecting network using compressed sensing for medical image segmentation. *Biomed. Signal Process. Control* **2023**, *86*, 105131. [[CrossRef](#)]
54. Garbaz, A.; Oukdach, Y.; Charfi, S.; El Ansari, M.; Koutti, L.; Salihoun, M. Bleeding Segmentation Based on a U-Formed Network with Separable Contextual Feature-Guided in Wireless Capsule Endoscopy Images. In Proceedings of the 2024 11th International Conference on Wireless Networks and Mobile Communications (WINCOM), Leeds, UK, 23–25 July 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1–6.
55. Feng, Y.; Su, J.; Zheng, J.; Zheng, Y.; Zhang, X. A parallelly contextual convolutional transformer for medical image segmentation. *Biomed. Signal Process. Control* **2024**, *98*, 106674. [[CrossRef](#)]
56. Wu, H.; Chen, S.; Chen, G.; Wang, W.; Lei, B.; Wen, Z. FAT-Net: Feature adaptive transformers for automated skin lesion segmentation. *Med. Image Anal.* **2022**, *76*, 102327. [[CrossRef](#)]
57. Wang, Y.; Zhang, Y.; Zhang, L.; Xu, Y.; Feng, R.; Cai, H.; Xue, J.; Zhao, Z.; Guo, X.; Wei, Y.; et al. Multi-Bottleneck progressive propulsion network for medical image semantic segmentation with integrated macro-micro dual-stage feature enhancement and refinement. *Expert Syst. Appl.* **2024**, *252*, 124179. [[CrossRef](#)]
58. Zhang, Y.; Liu, L.; Han, Z.; Meng, F.; Zhang, Y.; Zhao, Y. TranSEFusionNet: Deep fusion network for colorectal polyp segmentation. *Biomed. Signal Process. Control* **2023**, *86*, 105133. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.