applied sciences

MDPI

Article

A Fuzzy Transformer Fusion Network (FuzzyTransNet) for Medical Image Segmentation: The Case of Rectal Polyps and Skin Lesions

Ruihua Liu *, Siyu Duan, Lihang Xu, Lingkun Liu, Jinshuang Li and Yangyang Zou

School of Artificial Intelligence, Chongqing University of Technology, Chongqing 401135, China; duansy@stu.cqut.edu.cn (S.D.); 1052978252@stu.cqut.edu.cn (L.X.); liulingk@stu.cqut.edu.cn (L.L.); mettles@stu.cqut.edu.cn (J.L.); mathzyy@cqut.edu.cn (Y.Z.)
* Correspondence: lruih@cqut.edu.cn

Featured Application: Authors are encouraged to provide a concise description of the specific application or a potential application of the work. This section is not mandatory.

Abstract: Skin melanoma, one of the deadliest forms of cancer worldwide, demands precise diagnosis to mitigate cancer-related mortality. While histopathological examination, characterized by its cost-effectiveness and efficiency, remains the primary diagnostic approach, the development of an accurate detection system is pressing due to melanoma's varying sizes, shapes, and indistinct boundaries shared with normal tissues. To address the efficient segmentation of skin melanoma, we propose an innovative hybrid neural network approach in this study. Initially, a fuzzy neural network is constructed using fuzzy logic to preprocess medical images, supplemented by wavelet transformation for image enhancement. Subsequently, the Swin Transformer V2 and ResNet50 networks are introduced to parallelly extract features and apply them to the task of skin melanoma segmentation. Extensive experimental comparisons are conducted with other classic and advanced medical segmentation algorithms on publicly available skin datasets, namely ISIC 2017 and ISIC 2018. Experimental results reveal that our method outperforms the optimal algorithms by 1.3% in the Dice coefficient and 1.3% in accuracy on the ISIC 2018 dataset. The evaluation metrics indicate the effectiveness of the constructed fuzzy block in identifying uncertain lesion boundaries, while the Transformer–CNN branch adeptly extracts global features while accurately capturing underlying details. Additionally, we successfully apply our method to colon polyp segmentation tasks with similar indistinct boundaries, achieving remarkable segmentation outcomes.

Keywords: segmentation; medical image; fuzzy logic; CNN; Swin Transformer

check for updates

Citation: Liu, R.; Duan, S.; Xu, L.; Liu, L.; Li, J.; Zou, Y. A Fuzzy Transformer Fusion Network (FuzzyTransNet) for Medical Image Segmentation: The Case of Rectal Polyps and Skin Lesions. *Appl. Sci.* **2023**, *13*, 9121. https://doi.org/10.3390/app13169121

Academic Editor: Jan Egger

Received: 20 July 2023
Revised: 8 August 2023
Accepted: 9 August 2023
Published: 10 August 2023

Copyright: © 2023 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) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Melanoma is the most severe and fatal form of skin cancer [1], and dermatologists diagnose pigmented skin lesions and malignant melanomas through dermoscopy. While physicians currently employ computer-aided diagnosis (CAD) [2] and utilize robotics [3,4] to enhance surgical precision and accuracy for melanoma treatment, the segmentation of lesion regions often still necessitates manual labeling and review by experienced clinicians [5]. Furthermore, the intricate shapes and complexities of skin lesions (e.g., small sizes, intricate forms, and blurred boundaries due to color and texture) pose challenges for CAD systems based on traditional image segmentation algorithms, particularly in meeting the demands of precise image segmentation in the context of big data. Recently, deep learning-based approaches have exhibited significant advantages in enhancing segmentation accuracy over traditional methods based on thresholding, edge detection, and region- or graph-based techniques, particularly in the realms of medical and complex image segmentation.

Deep learning-based image segmentation methods can handle various types of image data. Among them, the U-Net convolutional neural network [6] has shown good performance in image segmentation. The network utilizes an encoder to extract features of different scales from the image and a decoder with skip connections to obtain a larger receptive field. Therefore, many subsequent medical image segmentation models have been implemented based on U-Net. Inspired by U-Net, UNet++ [7] introduces a series of nested dense convolutional blocks. This connection bridges the semantic gap between the encoder and decoder before feature fusion. In addition, new skip connection methods, such as residual connections [8] and dense connections [9], are also introduced into the network architecture. Although convolutional neural networks have become the mainstream method in medical image segmentation, these networks using convolutional operations inevitably have limitations in modeling long-range dependencies due to the local nature of inductive biases and weight sharing. Moreover, CNN-based segmentation methods still face difficulties in modeling and extracting global-level semantic features [10]. The emergence of Transformers can effectively address the aforementioned issues.

The Transformer is the first sequence-to-sequence prediction model to be used for NLP tasks [11]. Due to its ability to fully utilize the global information of input sequences and effectively model the global context, it has gradually gained widespread attention from computer vision researchers. Among them, the Swin Transformer [12] is one of the better-performing networks. The Swin Transformer addresses the computational and memory overhead issues of traditional Transformers when dealing with large-scale images by introducing a window-based attention mechanism. To overcome the limitations of local self-attention mechanisms in handling long-range dependencies, the Swin Transformer innovatively adopts the shift-window operation, dividing the image into multiple overlapping windows instead of uniformly partitioning it. Each window is processed as an independent image block, and there can be overlapping regions between windows. In this way, the Swin Transformer avoids performing global self-attention calculations on the entire image. Liu et al. [13] proposed residual afterglow normalization technology and the scaled cosine attention method based on the Swin Transformer, effectively transferring low-resolution pre-trained models under windows to corresponding models with higher resolutions. Additionally, in medical image processing tasks, to overcome the difficulties of convolutional networks in effectively extracting local features, the SwinE-Net segmentation network, proposed by Park et al. [14], maintains global semantics while fully utilizing the underlying features of the network.

Currently, many semantic segmentation tasks require high-resolution input images or large attention windows. However, the window sizes of the aforementioned Transformer-based networks can vary significantly between low-resolution pre-training and high-resolution fine-tuning. Currently, the most common approach is to perform bicubic interpolation on the position offset maps [12,15]. This simple fix often leads to suboptimal results. To address this issue, we adopt the Swin Transformer V2 for feature extraction and introduce the log-spaced continuous position bias method. It applies a small meta-network to logarithmically spaced coordinate inputs to generate bias values for arbitrary coordinate ranges. This method effectively transfers models pre-trained under low-resolution images and windows to corresponding models with higher resolutions. To maintain a strong grasp of the underlying details, the Swin Transformer V2 is combined with ResNet50, and their multi-level features are input into the fusion module. The fusion module first refines the multi-level feature maps, enhances their robustness and resolution, and then applies self-attention and bilinear Hadamard products to adaptively select fusion information. The fused feature maps are then combined, and segmentation images are generated using gated skip connections [16]. This approach enables more accurate and robust segmentation results for polyps and skin.

Moreover, the shape, location, and volume of skin melanoma vary among patients. The boundaries between lesion areas and surrounding normal tissues are often ambiguous, and significant variations in image color and contrast may arise due to disparities in

imaging equipment and settings employed by different healthcare institutions [17]. A new method, designed by Müller [18], is used to preprocess these images by skipping blank areas. It first identifies blank areas in the image through thresholding or other image processing techniques and marks these areas as regions that do not participate in training and segmentation. During training and inference, the model only processes non-blank areas. However, this method does not have a direct impact on dealing with blurred boundaries. In [19–22], the random displacement field method was adopted to achieve the smooth elastic deformation of medical images, increasing the diversity of image content. The application of the random displacement field method may increase image variation and texture to a certain extent, making some features of blurred boundaries more prominent and helping the model better adapt to situations with blurred boundaries. However, it should be noted that blurred boundaries are usually caused by the quality of the image itself or limitations in the imaging process, and the random displacement field method cannot be a reliable solution.

In response to the above issues, we propose a new preprocessing module that combines fuzzy logic. Fuzzy logic is a logical reasoning method for dealing with uncertainty and fuzziness and can be applied to image processing and segmentation tasks to solve the fuzzy boundary problem. Fuzzy logic classifies pixels into multiple fuzzy sets based on their grayscale values or other features and uses fuzzy rules to infer the degree of uncertainty of the boundary. In this way, fuzzy boundaries can be represented in a continuous and fuzzy manner, rather than just a binary boundary. In this paper, by using histogram equalization, wavelet transform, and the uncertainty of membership functions in fuzzy logic are utilized to address the medical image boundary fuzziness and imaging problems, it is applied to the input layer of the network to pre-label all pixels in the image, thereby improving the segmentation performance of the model on fuzzy boundaries.

Our main contributions are summarized as follows:

(1) In order to effectively deal with fuzzy boundaries, we propose a fuzzy block preprocessing method. This method uses membership functions to transform input features into the fuzzy domain and reduces the uncertainty of each pixel in the image through uncertainty mapping functions and fuzzy fusion functions, thereby improving the accuracy of the segmentation model.
(2) FuzzyTransNet is a new deep learning model that can be applied to polyp and skin segmentation. It accomplishes feature extraction by applying multi-correlated convolution, multi-feature aggregation, and attention mechanisms. Innovatively, it adopts the Swin Transformer V2 network for feature extraction and effectively combines it with ResNet50 through the fusion module. As a result, the network can extract global semantic information without sacrificing underlying contextual features.
(3) Experiments are conducted on three skin segmentation benchmark datasets (ISIC2017 and ISIC2018) and polyp segmentation datasets (Kvasir, CVC-ClinicDB, and CVC-ColonDB). Compared with some current advanced algorithms, our proposed method demonstrates advantages in evaluation metrics such as mDice, mIoU, and accuracy. Furthermore, by using class activation maps (Cams) for visualization, the effectiveness of the proposed algorithm is demonstrated more intuitively.

2. Methodology

As shown in Figure 1, FuzzyTransNet mainly consists of three parts: (1) A fuzzy preprocessing module in which we propose a new preprocessing approach to address the issues of blurry boundaries and imaging problems in medical images. This approach is applied to the input layer of the network. (2) ResNet50 and Swin Transformer V2 are chosen as the two branches for obtaining local and global features. The multi-level features extracted from both branches are then fed into the fusion module, allowing for effective integration. (3) An attention-gated (AG) network [16] and skip connection are used to combine the features output from each fusion module to generate the final segmentation result.

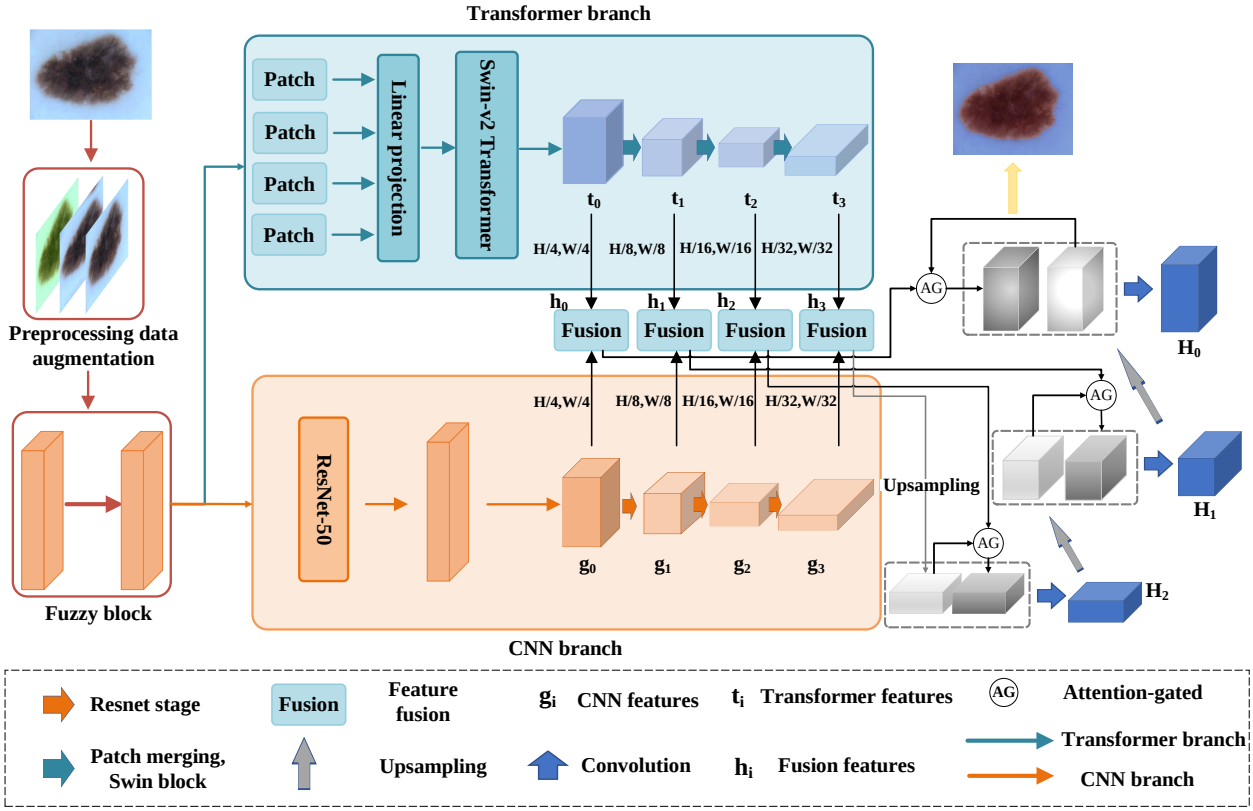


Figure 1. Pipeline of the proposed segmentation framework.

2.1. Fuzzy Block

Fuzzy logic has the ability to express the uncertainty of things and make decisions through fuzzy processing and operations, and it can be effectively combined with neural networks to produce good results. Because the Gaussian function has the capacity for non-linearity and fitting functions, the fuzzy block constructed in this paper uses the Gaussian function as the membership function within the neural network, as shown in the following formula:

$$\mu(x) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} exp\left(-\frac{1}{2}x^T\Sigma^{-1}x\right) \quad (1)$$

The fuzzy block structure is shown in Figure 2. In the fuzzy layer, a feature map is input with dimensions (N, H, W, C), and the number of original image categories is M. In order to calculate the Gaussian function, the two-dimensional pixel value in the feature map needs to be changed to one-dimensional data (H × W) through the Gaussian membership function to calculate the degree of certainty that each pixel value belongs to the lesion area. Since the output value of the Gaussian function varies over a wide range, the L2 regularization method is used to process it so that it can be normalized to a value range similar to the distribution of the original input data. After the fuzzy layer outputs the fuzzy logic uncertainty value with dimensions (N, H, W, M), it can affect the distribution of the pixel value information in the feature map. Finally, the uncertainty value and the original input are fused into a feature map with dimensions (N, H, W, C + M) and then output to the neural network for subsequent operations.

Fuzzy layer: If the membership degree of pixel *i* is close to 1 or 0, it means that the pixel is likely to be a lesion or background, which implies a high degree of certainty. If the membership degree of pixel *i* is close to 0.5, it is difficult to determine whether the pixel

belongs to the lesion area or the background. $\mu(x_i)$ is the membership degree of pixel i, and Formula (2) is used to calculate its uncertainty and obtain the uncertainty $o(x_i)$ of pixel i

$$o(x_i) = \begin{cases} 2 \times \mu(x_i) & \text{if } \mu(x_i) < 0.5 \\ 2 \times (1 - \mu(x_i)) & \text{if } \mu(x_i) \geq 0.5 \end{cases} \tag{2}$$

Merge: By fusing the uncertainty of the pixels with the original input features, the effect of reducing the uncertainty of the blurred boundaries of the original features is achieved, as shown in Formula (3).

$$p_i = (1 - o(x_i)) \cdot x_i \tag{3}$$

where $o(x_i)$ is the fuzzy membership degree of pixel i obtained using Formula (1), and x_i is the original feature of the pixel. This equation shows that if the uncertainty of a pixel is closer to 1, its weight should be decreased. Finally, the obtained result p_i is output to the neural network for subsequent operations.

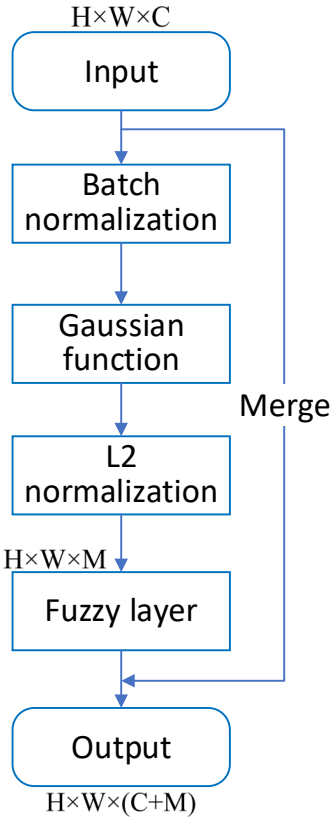


Figure 2. Fuzzy block structure visualization.

2.2. Transformer Branch

The design of the Transformer branch follows the codec structure. As shown in Figure 1, an image with a size of $x \in R^{H \times W \times C}$ is divided into $N = \frac{H}{S} \times \frac{W}{S}$ patch blocks of equal size in the Transformer branch, where S is usually set to 16. After the patch blocks are divided, they are flattened and passed to the linear embedding layer with an output dimension of D_0, and the original embedding sequence $e \in R^{N \times D_0}$ is obtained. To exploit spatial prior information, a learnable positional embedding of the same dimension is added under e to generate e with positional embeddings $z^0 \in R^{N \times D_0}$. The Swin Transformer improves the attention mechanism, as shown in the following formula:

$$Attention(Q, K, V) = SoftMax\left(QK^T / \sqrt{d} + B\right)V \tag{4}$$

where $Q, K, V \in R^{M^2 \times d}$ are the query, key, and value matrices, and d is the dimension. $B \in R^{M^2 \times M^2}$ is the relative position bias, which determines the relative position of each token in the coding window, M is the size of the window, and there are M^2 tokens in total. By defining a training bias matrix $\hat{B} \in R^{(2M-1)\times(2M-1)}$, B can be obtained through indexing.

The improvement strategy of the Swin Transformer V2 involves using continuous relative position bias, no longer defining a fixed size, but using a small network to predict relative position bias, as shown in the following formula:

$$B(\Delta x, \Delta y) = G(\Delta x, \Delta y) \tag{5}$$

where G is an MLP model containing 2 layers with ReLU activation in the middle. The advantage of using network G is that it can generate relative position bias at any relative position so that no changes are required when migrating to a larger window. Although network G can adapt to windows of different sizes, when the window size changes, the relative position range also changes, which means that network G has to accept a different input range than the pre-trained model. In order to minimize the change in the input range, we further propose replacing the original linear coordinates with the coordinates in the log space. The conversion formula of the two is as follows:

$$\begin{aligned} \widehat{\Delta x} &= sign(x) \cdot \log(1 + |\Delta x|) \\ \widehat{\Delta y} &= sign(y) \cdot \log(1 + |\Delta y|) \end{aligned} \tag{6}$$

where Δx, Δy are the linear-scaled coordinates and $\widehat{\Delta x}$, $\widehat{\Delta y}$ are the log-spaced coordinates.

2.3. CNN Branch

In this paper, an image with a size of $x \in R^{H \times W \times C}$ is input into a CNN feature extraction network composed of ResNet50, and the Transformer branch is utilized to capture global contextual information. The ResNet-based model typically consists of five blocks, with each block performing downsampling of the feature maps by a factor of two. The first four blocks produce feature maps denoted as $g_3 \in R^{\frac{H}{32} \times \frac{W}{32} \times C_3}$, $g_2 \in R^{\frac{H}{16} \times \frac{W}{16} \times C_2}$, $g_1 \in R^{\frac{H}{8} \times \frac{W}{8} \times C_1}$, and $g_0 \in R^{\frac{H}{4} \times \frac{W}{4} \times C_0}$, respectively, where H, W, and C_i represent the height, width, and number of channels of the feature map denoted as g_i. These feature maps are then fed into the fusion block, where they are combined with the results from the Transformer branch to obtain the final interactive features (Figure 1).

2.4. Fusion Block

In order to effectively integrate the encoding features from both the CNN branch and Transformer branch, the features are fused using the fusion block, which combines the channel attention mechanism and the spatial attention mechanism. Specifically, the fused feature representation is obtained through the following operations.

$$T_i = ChannelAttn(t_i) \tag{7}$$

$$G_i = SpatialAttn(g_i) \tag{8}$$

$$b_i = Conv\left(t_i W_1^i \odot g_i W_2^i\right) \tag{9}$$

$$h_i = Residual([b_i, T_i, G_i]) \tag{10}$$

where $W_1^i \in {}^{D_i \times L_i}$, $W_2^i \in {}^{C_i \times L_i}$, $\odot|$ are the Hadamard products, and Conv is a 3×3 convolutional layer.

Channel attention is implemented using the SE Block proposed in [23], as shown in Figure 3, to facilitate the incorporation of global information from the Transformer branch. In the fusion block, channel attention is applied first to identify the specific positions of globally significant information in the feature maps from the Transformer

branch, aiming to enhance the feature representation capability. Due to the presence of noise in the raw data captured by image sensors and the aggregation of operations such as convolution and pooling during feature extraction, this noise may be further amplified. Consequently, in the CNN branch, low-level features may be affected by certain levels of noise. Ref. [24] proposed weighting in the spatial dimension of the feature maps. Through spatial attention, the network can better focus on crucial regions while disregarding irrelevant areas. Subsequently, the feature matrices from the two branches, with the same dimensions, are element-wise multiplied using the Hadamard product. Next, the concatenated interaction feature b_i and attention features G_i, T_i are computed using the residual block. The resulting feature h_i effectively captures the global and local context of the current spatial resolution. Finally, the attention-gated (AG) method [16] and skip connection are combined (see Equation (11)) to generate the final segmentation result, as shown in Figure 1.

$$H_{i-1} = Conv([Up(H_i), AG(h_{i-1}, Up(H_i))]) \tag{11}$$

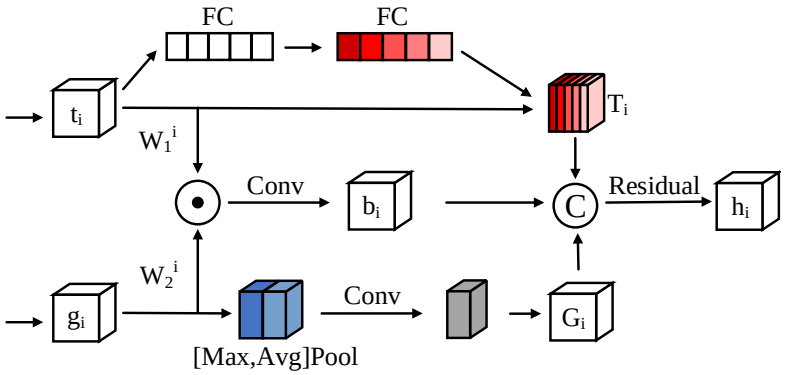


Figure 3. Fusion module structure.

3. Skin Lesion Experiments and Analysis

In this study, we conduct experiments on publicly available skin lesion datasets, namely ISIC 2017 [25] and ISIC 2018 [26], and compare and evaluate our model against other state-of-the-art algorithms. Subsequently, the datasets are introduced, followed by a comparative evaluation of our model with other advanced algorithms in terms of experimental metrics. Finally, ablation experiments are performed to validate the effectiveness of the various modules in the model.

The experimental setup in this study involved an AMD Ryzen 5600X@3.70 GHz CPU and a single NVIDIA GeForce RTX 3080 ti GPU with 12 GB of memory and a total of 32 GB RAM. The experiments were conducted using the PyTorch framework, and the model optimization employed the mini-batch stochastic gradient descent method during the training process. Each experiment consisted of 70 epochs, with a training batch size of 4, a learning rate of 7×10^{-5}, and a momentum parameter of 0.9.

3.1. Datasets

The International Skin Imaging Collaboration (ISIC) is an international project aimed at facilitating the diagnosis and research of skin lesions. The ISIC dataset provides a substantial collection of skin lesion images, including examples of nodular melanoma, freckled melanoma, and superficial spreading melanoma. Notably, two pivotal datasets within the ISIC project are ISIC 2017 [25] and ISIC 2018 [26], which are employed for skin lesion classification and segmentation studies. ISIC 2017 consists of 2000 training images and 600 melanoma testing images, serving as the evaluation dataset for the proposed approach. Likewise, ISIC 2018 comprises 1816 training images and 778 melanoma testing images, establishing itself as a primary benchmark for medical image algorithm evaluation.

In this study, both ISIC datasets were resampled to 224 × 320 pixels, with ISIC 2018 being partitioned into a training set (70%), validation set (10%), and testing set (20%).

3.2. Evaluation Criteria

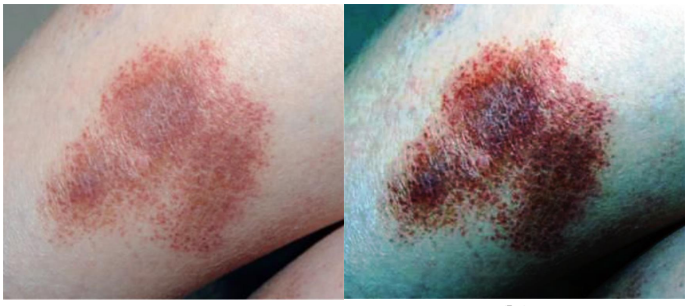
In this study, the performance of FuzzyTransNet and various models on the aforementioned datasets was evaluated using four widely used evaluation criteria, including Dice, mIoU, the Jaccard Index (JI), and accuracy (ACC). The specific details are shown in Table 1. X and Y represent the ground truth and the predicted result. *TP*, *TN*, *FP*, and *FN* are True Positive, True Negative, False Positive, and False Negative, respectively.

Table 1. Evaluation metrics.

Metric	Description
Dice	$\frac{2\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert}$
mIoU	$\frac{1}{k+1}\sum_{i=0}^{k}\frac{TP}{FN+FP+TP}$
Jaccard Index (JI)	$\frac{TP}{TP+FN+FP}$
Accuracy (ACC)	$\frac{TP+TN}{TP+TN+FP+FN}$

3.3. Preprocessing Data Augmentation

Initially, a preprocessing enhancement method was applied to the images prior to entering the fuzzy block. The original grayscale image with three identical channels underwent three separate treatments: histogram equalization, extraction of low-frequency information through wavelet transform, and extraction of high-frequency information through wavelet transform. Histogram equalization retained the core information of the original grayscale image, while the extraction of low-frequency information via wavelet transform enhanced image contrast and the extraction of high-frequency information highlighted detailed features across different regions of the image. As depicted in Figure 4, this image-processing approach resulted in a more distinct demarcation between lesion areas and the background, effectively alleviating the issues of low contrast and noise prevalent in medical images.



(a) (b)

Figure 4. (**a**) Original image. (**b**) Image after histogram equalization and wavelet transformation.

3.4. Results of Skin Lesion Segmentation

The loss function of the FuzzyTransNet proposed in this paper consists of a binary cross-entropy loss employing the sigmoid activation function. In this paper, the evaluation metrics used on the ISIC dataset include the Dice coefficient, Jaccard Index (JI), and pixel-wise accuracy (ACC). A comparison with state-of-the-art methods on the ISIC2017 dataset is presented in Table 2. As shown in Table 2, our proposed algorithm achieved superior performance in terms of the JI, Dice coefficient, and ACC, surpassing U-Net by 3.6%, 3.9%, and 2.7%, respectively. It also outperformed DCL-PSI by 2.2%, 3.1%, and 1.4%, and Ms RED by 1.4%, 0.7%, and 1.4%. On the ISIC2018 dataset, which contains more complex

skin lesions, the lesion boundaries are more ambiguous compared to the ISIC2016 and ISIC2017 datasets, making it challenging to separate the lesion area from the background. Consequently, most models exhibited lower performance on the ISIC2018 dataset compared to the ISIC2016 and ISIC2017 datasets. For instance, U-Net++ achieved metrics of 76.1%, 85.1%, and 92.5% on the ISIC2018 dataset, reflecting a decrease of 1.2%, 0.4%, and 1.3% compared to the ISIC2017 dataset. In contrast, FuzzyTransNet demonstrated favorable results in terms of the JI, Dice coefficient, and ACC, reaching 80.1%, 88.9%, and 96.1%, as shown in Table 3. In Table 3, it is evident that our proposed method outperformed the previous state-of-the-art approach, CA-Net, by 0.7%, 2.5%, and 1.3% in terms of the JI, Dice coefficient, and ACC. Furthermore, compared to using ResNet34 and ViT as the backbone in TransFuse, FuzzyTransNet exhibited an improvement of 1.0%, 2.0%, and 1.9%. However, on the ISIC2017 dataset, FuzzyTransNet only slightly surpassed TransFuse by 0.4%, 1.4%, and 1.1%. These findings highlight the greater advantages of FuzzyTransNet in the face of more complex lesion boundaries in medical images. This can be attributed to the well-designed data augmentation block and fuzzy block, which effectively preprocessed the experimental data.

Table 2. Quantitative results on the ISIC2017 test set.

Method	Year	JI	Dice	ACC
U-Net [6]	2015	0.763	0.847	0.928
CDNN [27]	2017	0.765	0.849	0.934
DDN [28]	2018	0.765	0.866	0.940
FrCN [29]	2018	0.771	0.871	0.940
SLSDeep [30]	2018	0.782	0.878	0.936
DCL-PSI [31]	2019	0.777	0.857	0.941
U-Net++ [7]	2019	0.773	0.855	0.938
TransFuse [32]	2021	0.795	0.872	0.944
Ms RED [33]	2022	0.786	0.865	0.941
Our method	-	**0.799**	**0.886**	**0.955**

Table 3. Quantitative results on the ISIC2018 test set.

Method	Year	JI	Dice	ACC
U-Net [6]	2015	0.755	0.824	0.895
U-net++ [7]	2019	0.761	0.851	0.925
DCL-PSI [31]	2019	0.763	0.847	0.929
DO-Net [34]	2020	0.794	0.854	0.941
CA-Net [35]	2020	0.794	0.874	0.948
TransFuse [32]	2021	0.791	0.869	0.942
Ms RED [33]	2022	0.793	0.876	0.946
Our method	-	**0.801**	**0.889**	**0.961**

3.5. Ablation Study

To assess the effectiveness of FuzzyTransNet in the context of melanoma skin lesion segmentation, we undertook an evaluation that encompassed various Transformer branches and CNN branches. This evaluation was conducted through the utilization of both sequential and parallel connections, employing the Dice coefficient as our benchmark metric. As demonstrated in Table 4, when comparing the results of E.1, E.2, and E.3, we can discern a marginal performance gain with ResNet50 over ResNet34. Furthermore, it is noteworthy that the Swin Transformer V2 outperformed DeiT-Small [36] in terms of performance. When juxtaposing E.3 with E.6, it becomes evident that the parallel models exhibited superior performance over the sequential models. Moreover, when contrasting E.6 with E.4 and E.5, we can see that the Swin Transformer V2, in contrast to the other Transformers, effectively translated low-resolution images and pre-trained models from a windowed perspective to higher resolutions, thereby further amplifying the overall performance. Additionally, when comparing the results of E.4, E.6, and E.7, we can glean

that with a mere replacement of dual branches, FuzzyTransNet's performance surpassed that of TransFuse. Notably, with the integration of the fuzzy block, the Dice coefficient was significantly enhanced. Within the context of the ISIC2017 and ISIC2018 datasets, E.7 exhibited increments of 0.7% and 0.8%, respectively, compared to E.6.

Additional experiments conducted on the ISIC2018 dataset are presented in Table 5. After we used ResNet50 and the Swin Transformer V2 as the two branches, the three evaluation indicators increased by 0.6%, 1.1%, and 0.9%, respectively, compared to ResNet34 and DeiT-S, which were used by TransFuse. After introducing the fuzzy block constructed in this paper, the result after changing the branch was further improved by 0.4%, 0.8%, and 0.7%. Finally, the data augmentation model was added to achieve the optimal result.

Table 4. Ablation study on parallel-in-branch design. R34: ResNet34; R50: ResNet50; Swin V1: Swin Transformer; Swin V2: Swin Transformer V2; DeiT-S: DeiT-Small.

Index	Backbones	Sequential	Parallel	Fusion	ISIC2017	ISIC2018
E.1	R34 + DeiT-S	✓	-	-	0.849	0.852
E.2	R50 + DeiT-S	✓	-	-	0.853	0.855
E.3	R50 + Swin V2	✓	-	-	0.859	0.862
E.4	R34 + DeiT-S	-	✓	✓	0.872	0.869
E.5	R50 + Swin V1	-	✓	✓	0.874	0.875
E.6	R50 + Swin V2	-	✓	✓	0.877	0.879
E.7	E.6 + Fuzzy block	-	✓	✓	**0.884**	**0.887**

Table 5. Component-wise analysis.

Index	R34 + DeiT-S	R50 + Swin V2	Fuzzy Block	Data Augmentation	Jaccard	Dice	Accuracy
TransFuse	✓	-	-	-	0.791	0.868	0.942
F.2	-	✓	-	-	0.797	0.879	0.951
F.3	-	✓	✓	-	0.801	0.887	0.958
Ours	-	✓	✓	✓	**0.801**	**0.889**	**0.961**

3.6. Visual Comparison

To further assess the effectiveness and robustness of our proposed FuzzyTransNet, this study conducted additional visual comparison experiments on representative images from the ISIC2018 dataset. Figure 5 presents visual comparisons of the segmentation results on three typical skin samples using the classic algorithms U-Net [6], U-Net++ [7], and TransFuse [32], as well as the proposed FuzzyTransNet. When compared to the ground truth, our algorithm's segmentation results exhibit greater similarity to the ground-truth images compared to the other algorithms. In the second row of Figure 5, it is evident that the three classic models exhibit more errors in segmenting fuzzy lesion boundaries. Conversely, the FuzzyTransNet model results in a clearer and more complete contour, underscoring the effectiveness of incorporating the fuzzy block for processing.

To further substantiate the efficacy of our proposed algorithm, we employed the Grad-CAM (gradient-weighted class activation mapping) [37] technique for visualizing the class activation maps. Concurrently, we conducted a comparative analysis with representative medical image segmentation algorithms (U-Net [6], U-Net++ [7], and TransFuse [32]) to vividly demonstrate the advantages of our approach. The visual comparative results are presented in Figure 6. In the heatmaps in the first row, it can be seen that FuzzyTransNet exhibits a superior focus on lesion areas compared to the other models. In the second row of the segmentation results, it is evident that TransFuse [32] demonstrates less distinct attention to lesion boundaries, resulting in a less clear delineation of the lesion region. Conversely, our proposed method exhibits a more explicit emphasis on lesion boundaries, thus indirectly corroborating the effectiveness of the fuzzy block introduced in this study.

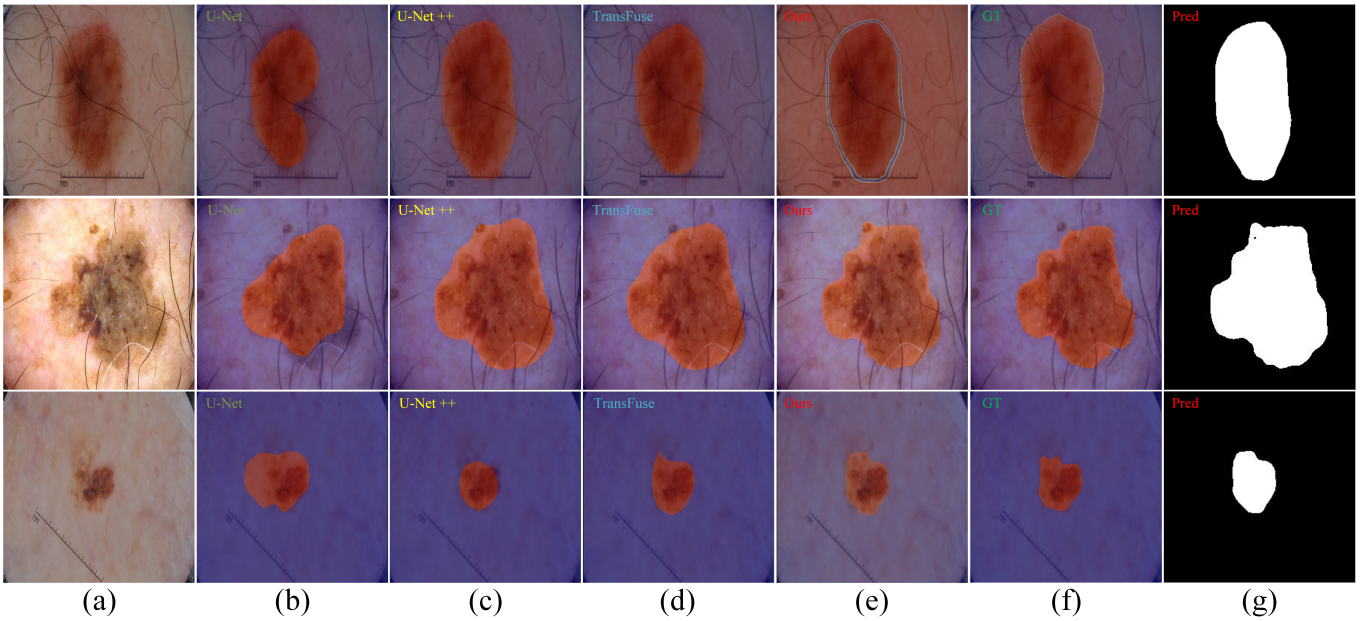


Figure 5. Comparison of representative skin lesion segmentation results. (**a**) Original image. (**b**) Results of U-Net. (**c**) Results of U-Net++. (**d**) Results of TransFuse. (**e**) Results of **our method**. (**f**) Ground truth. (**g**) Segmentation results generated by **our method** (grayscale).

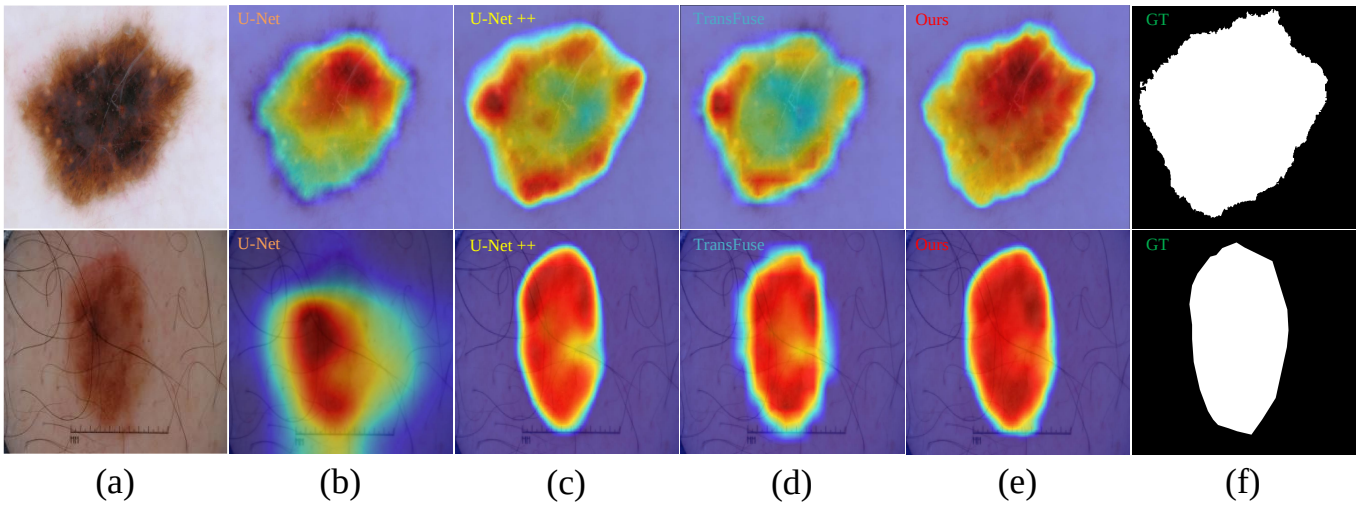


Figure 6. Comparison of the CAMs obtained using the representative algorithms. (**a**) Original image. (**b**) CAMs generated by U-Net. (**c**) CAMs generated by U-Net++. (**d**) CAMs generated by TransFuse. (**e**) CAMs generated by **our method**. (**f**) Ground truth.

4. Model Applied to Polyp Segmentation

Although in the preceding sections, FuzzyTransNet achieved satisfactory results on the ISIC 2017 and ISIC 2018 datasets, we aim to demonstrate its remarkable generalization capability and effectiveness in a broader scope of medical image segmentation, encompassing diverse types of lesion scenarios. Given the resemblance between the rectal polyp dataset and the skin lesion dataset in certain aspects, such as high-resolution training images and the lack of distinct boundaries between lesion regions and surrounding normal tissues, in this section, we apply our model to three rectal polyp datasets, namely CVC-ColonDB [38], Kvasir [39], and CVC-linicDB [40], to validate its efficacy.

4.1. Datasets

CVC-ColonDB [38]: This dataset comprises 300 images collected from 15 colonoscopy video sequences, with a resolution of 288 × 384. Kvasir [39] is an open dataset of gastrointestinal polyp images and corresponding segmentation masks, manually annotated and validated by experienced gastroenterologists. It includes 1000 polyp images from Kvasir Dataset v2, along with their corresponding segmentation masks, with images stored using JPEG compression encoding. The images in Kvasir-SEG vary in resolution from 332 × 487 to 1920 × 1072 pixels, thus necessitating uniform resizing. In our experiments, the image size was set to 512 × 512. CVC-linicDB [40]: This dataset consists of 612 images with dimensions of 500 × 570. Ground-truth images for the aforementioned datasets are provided alongside the data and compiled by a team of medical professionals, serving as the benchmark for the classification and segmentation outcomes.

4.2. Results of Polyp Segmentation

Firstly, an evaluation of the proposed segmentation method in this paper was conducted based on the Dice and mIoU evaluation metrics. Table 6 quantitatively presents the performance of FuzzyTransNet and eight comparative methods on three colon polyp datasets. All evaluation metrics in the table are the averaged results of a sixfold cross-validation. It can be observed that FuzzyTransNet achieved favorable performance on the Kvasir and ClinicDB datasets. On the Kvasir dataset, FuzzyTransNet achieved Dice and mIoU metrics of 92.9% and 87.5%, respectively, which were 1.1% and 0.7% higher than those of TransFuse. On the ClinicDB dataset, FuzzyTransNet achieved Dice and mIoU metrics that were 0.8% and 0.5% higher than those of TransFuse. It is noteworthy that TransFuse performed the best among the eight segmentation networks, as seen in the table. However, FuzzyTransNet did not achieve the optimal performance on the ColonDB dataset. This can be attributed to the fact that the ColonDB dataset consists of 300 images collected from 15 colonoscopy video sequences, leading to a relatively limited amount of effective data due to the presence of multiple similar images. Consequently, while all segmentation networks exhibited a noticeable decrease in performance on the ColonDB dataset compared to the Kvasir and ClinicDB datasets, there was an improvement in the Dice coefficient for FuzzyTransNet compared to TransFuse.

Table 6. Quantitative results on polyp segmentation datasets of our method compared to other segmentation algorithms. '-' means results not available.

Method	Year	Kvasir		ClinicDB		ColonDB	
		Dice	mIoU	Dice	mIoU	Dice	mIoU
U-Net [6]	2015	0.818	0.746	0.823	0.750	0.512	0.444
U-Net++ [7]	2019	0.821	0.743	0.794	0.729	0.483	0.410
ResUNet++ [41]	2019	0.813	0.793	0.796	0.796	-	-
PraNet [42]	2020	0.813	0.793	0.796	0.796	-	-
HarDNetMSEG [43]	2021	0.912	0.857	0.932	0.882	0.731	0.660
TransUnet [44]	2021	0.913	0.857	0.932	0.883	**0.780**	**0.697**
TransFuse [32]	2021	0.918	0.868	0.934	0.886	0.773	0.696
Ms RED [33]	2022	0.911	0.854	0.934	0.885	0.779	0.690
Our method	-	**0.929**	**0.875**	**0.942**	**0.891**	0.778	0.689

4.3. Ablation Study

We also evaluated the effectiveness of FuzzyTransNet in colon polyp segmentation through an ablation study. In Table 7, when comparing F.1 and F.2, we can observe that the Swin Transformer V2 outperformed DeiT-Small [36] with the same CNN branch configuration. Furthermore, when comparing F.3 and F.4, we note that the parallel model demonstrated superior performance over the sequential model, underscoring the efficacy of our fusion module. Additionally, when comparing F.4 and F.5, we can observe a significant

improvement in the Dice coefficient after the incorporation of the fuzzy block. F.5 achieved a 0.6% improvement over F.4 on both the Kvasir and ClinicDB datasets.

Table 7. Ablation study on parallel-in-branch design. R34: ResNet34; R50: ResNet50; Swin V2: Swin Transformer V2; DeiT-S: DeiT-Small.

Index	Backbones	Sequential	Parallel	Fusion	Kvasir	ClinicDB
F.1	R50 + DeiT-S	✓	-	-	0.909	0.919
F.2	R50 + Swin V2	✓	-	-	0.911	0.925
F.3	R34 + DeiT-S	✓	-	-	0.918	0.934
F.4	R34 + DeiT-S	-	✓	✓	0.921	0.936
F.5	R50 + Swin V2	-	✓	✓	**0.927**	**0.942**

4.4. Visual Comparison

To further evaluate the effectiveness and robustness of our proposed FuzzyTransNet method, this paper conducted visual comparative experiments on representative images from both polyp and skin lesion datasets. Given the frequent oversight of polyps in routine colonoscopy examinations, with miss rates ranging from 14% to 30%, we selected two highly representative polyp samples closely adhering to tissue surfaces in Figure 7. In the original images, the color of polyps is often indistinct from that of the surrounding tissues. By observing the results in the second row, it can be stated that the segmentation outcomes of our proposed algorithm exhibit greater similarity to the ground truth compared to the other algorithms. Moreover, we employed the GradCAM (gradient-weighted class activation mapping) [37] visualization method to intuitively discern where each model concentrates its attention in lesion areas, particularly for polyps that may dangle or adhere closely to tissue surfaces. In Figure 8, it is evident that FuzzyTransNet focuses more comprehensively on lesion regions compared to the other models.

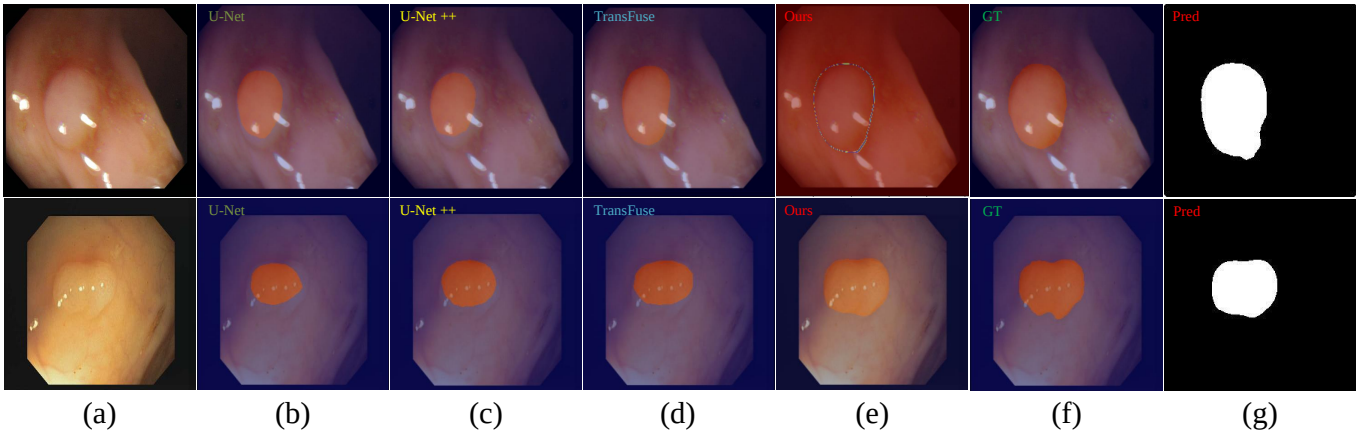


Figure 7. Comparison of representative polyp segmentation results. (**a**) Original image. (**b**) Results of U-Net. (**c**) Results of U-Net++. (**d**) Results of TransFuse. (**e**) Results of **our method**. (**f**) Ground truth. (**g**) Segmentation results generated by **our method** (grayscale).

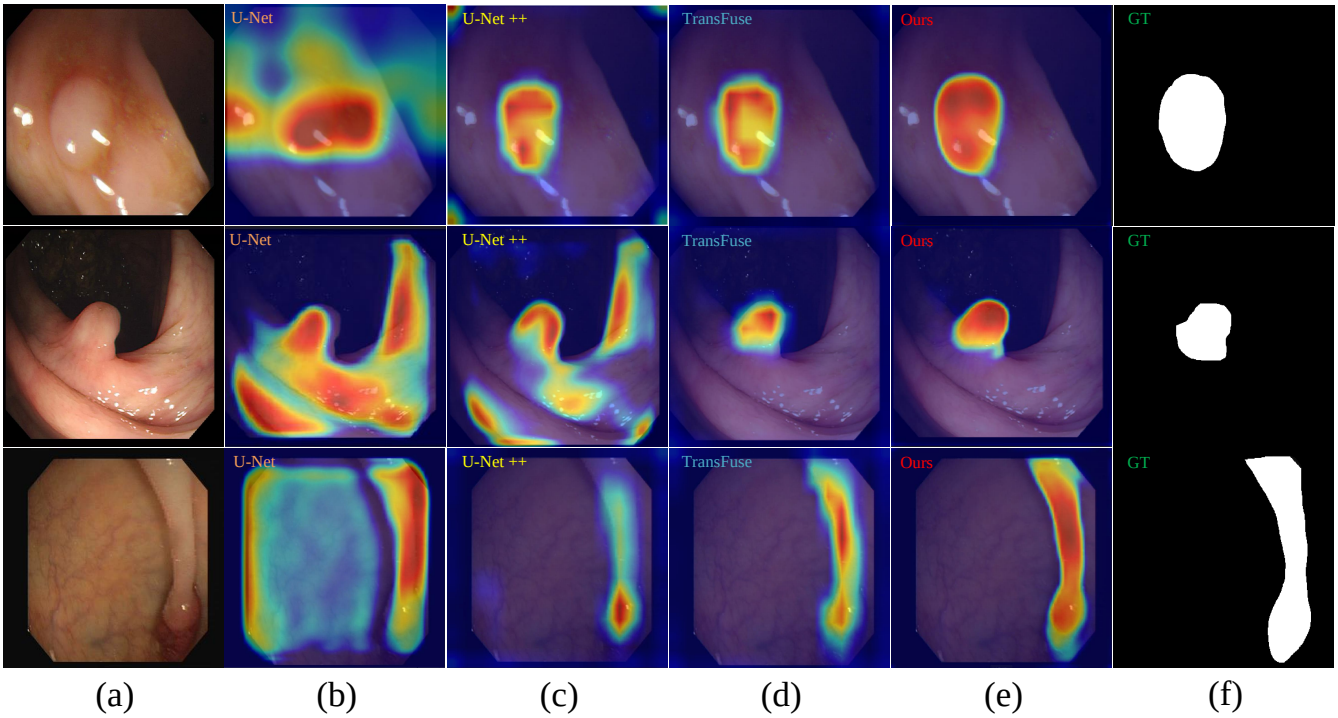


Figure 8. Comparison of the CAMs obtained using the representative algorithms. (**a**) Original image. (**b**) CAMs generated by U-Net. (**c**) CAMs generated by U-Net++. (**d**) CAMs generated by TransFuse. (**e**) CAMs generated by **our method**. (**f**) Ground truth (grayscale).

5. Conclusions

This paper proposes a new method based on medical image segmentation, achieving good segmentation results in rectal polyp and skin melanoma segmentation tasks. The method comprises three steps. First, in order to obtain more essential image features, we use contrast enhancement and wavelet transform to initially process the input image. Next, a fuzzy block is creatively used to blur the input image through a fuzzy membership function to suppress irrelevant regions, and it is fused with the original image to generate an uncertainty map as input. Finally, the Swin Transformer V2 and ResNet50 networks are introduced to extract image features, and channel attention and spatial attention are used to enhance global key information and local details, respectively. Experimental results on public datasets, such as Kvasir, ClinicDB, ISIC 2017, and ISIC 2018, show that the method proposed in this paper is superior to other models in terms of accuracy. In particular, the method exhibits significant improvements and robustness when dealing with small melanin regions and blurred polyp boundaries.

This study primarily focuses on addressing the challenge of fuzzy boundary segmentation in tasks involving skin lesions and polyp segmentation. We aim to extract comprehensive global semantic information without compromising the underlying contextual features. However, it is important to acknowledge that our proposed method has certain limitations. For example, it demonstrates significant advantages over other techniques mainly when dealing with datasets of specific resolutions and cases where boundaries are less distinct. Consequently, our future research endeavors will concentrate on refining our model by enhancing the feature extraction mechanisms for challenging regions and advancing methods for contextual information retrieval. Our goal is to transform our approach into a versatile medical image segmentation model capable of accommodating a broad spectrum of lesion types.

Author Contributions: Conceptualization, R.L. and S.D.; methodology, R.L. and S.D.; software, S.D.; validation, L.X., L.L. and J.L.; formal analysis, R.L. and J.L.; investigation, L.X.; resources, L.L.; data curation, S.D. and L.X.; writing—original draft preparation, S.D.; writing—review and editing, R.L., S.D., L.X., L.L. and Y.Z.; visualization, S.D.; supervision, R.L. and Y.Z.; project administration, R.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Chongqing Natural Science Foundation Project (Grant No. CSTB2023NSCQ-MSX0319) and the Science and Technology Project of Chongqing Municipal Education Commission (Grant No. KJQN202001129).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

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

References

1. Xia , C.; Dong, X.; Li, H.; Cao, M.; Sun, D.; He, S.; Yang, F.; Yan, X.; Zhang, S.; Li, N.; et al. Cancer statistics in China and United States, 2022: Profiles, trends, and determinants. *Chin. Med. J.* **2022**, *135*, 584–590. [CrossRef] [PubMed]
2. Hassan, C.; Wallace, M.B.; Sharma, P.; Maselli, R.; Craviotto, V.; Spadaccini, M.; Repici, A. New artificial intelligence system: First validation study versus experienced endoscopists for colorectal polyp detection. *Gut* **2020**, *69*, 799–800. [CrossRef]
3. Kim, Y.; Genevriere, E.; Harker, P.; Choe, J.; Balicki, M.; Regenhardt, R.W.; Vranic, J.E.; Dmytriw, A.A.; Patel, A.B.; Zhao, X. Telerobotic neurovascular interventions with magnetic manipulation. *Sci. Robot.* **2022**, *7*, eabg9907. [CrossRef] [PubMed]
4. Jin, D.; Wang, Q.; Chan, K.F.; Xia, N.; Yang, H.; Wang, Q.; Yu, S.C.H.; Zhang, L. Swarming self-adhesive microgels enabled aneurysm on-demand embolization in physiological blood flow. *Sci. Adv.* **2023**, *9*, eadf9278. [CrossRef] [PubMed]
5. Pittiglio, G.; Chandler, J.H.; da Veiga, T.; Koszowska, Z.; Brockdorff, M.; Lloyd, P.; Barry, K.L.; Harris, R.A.; McLaughlan, J.; Pompili, C.; et al. Personalized magnetic tentacles for targeted photothermal cancer therapy in peripheral lungs. *Commun. Eng.* **2023**, *2*, 50. [CrossRef]
6. 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; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
7. 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]
8. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
9. Li, X.; Chen, H.; Qi, X.; Dou, Q.; Fu, C.W.; Heng, P.A. H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Trans. Med. Imaging* **2018**, *37*, 2663–2674. [CrossRef]
10. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; Timofte, R. Swinir: Image restoration using swin transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 10–17 October 2021; pp. 1833–1844.
11. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
12. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 10–17 October 2021; pp. 10012–10022.
13. Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. Swin transformer v2: Scaling up capacity and resolution. In Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12009–12019.
14. Park, K.B.; Lee, J.Y. SwinE-Net: Hybrid deep learning approach to novel polyp segmentation using convolutional neural network and Swin Transformer. *J. Comput. Des. Eng.* **2022**, *9*, 616–632.
15. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
16. Schlemper, J.; Oktay, O.; Schaap, M.; Heinrich, M.; Kainz, B.; Glocker, B.; Rueckert, D. Attention gated networks: Learning to leverage salient regions in medical images. *Med. Image Anal.* **2019**, *53*, 197–207. [CrossRef]
17. Wang, T.; Ugurlu, H.; Yan, Y.; Li, M.; Li, M.; Wild, A.M.; Yildiz, E.; Schneider, M.; Sheehan, D.; Hu, W.; et al. Adaptive wireless millirobotic locomotion into distal vasculature. *Nat. Commun.* **2022**, *13*, 4465. [CrossRef]
18. Müller, D.; Kramer, F. MIScnn: A framework for medical image segmentation with convolutional neural networks and deep learning. *BMC Med. Imaging* **2021**, *21*, 12. [CrossRef]

19. Javaid, U.; Dasnoy, D.; Lee, J.A. Semantic segmentation of computed tomography for radiotherapy with deep learning: Compensating insufficient annotation quality using contour augmentation. In Proceedings of the Medical Imaging 2019: Image Processing, San Diego, CA, USA, 19–21 February 2019; SPIE: Bellingham, WA, USA, 2019; Volume 10949, pp. 682–694.
20. Lorenzo, P.R.; Nalepa, J.; Bobek-Billewicz, B.; Wawrzyniak, P.; Mrukwa, G.; Kawulok, M.; Ulrych, P.; Hayball, M.P. Segmenting brain tumors from FLAIR MRI using fully convolutional neural networks. *Comput. Methods Programs Biomed.* **2019**, *176*, 135–148. [CrossRef]
21. Wang, Y.; Li, C.; Zhu, T.; Zhang, J. Multimodal brain tumor image segmentation using WRN-PPNet. *Comput. Med. Imaging Graph.* **2019**, *75*, 56–65. [CrossRef]
22. Karani, N.; Erdil, E.; Chaitanya, K.; Konukoglu, E. Test-time adaptable neural networks for robust medical image segmentation. *Med. Image Anal.* **2021**, *68*, 101907. [CrossRef] [PubMed]
23. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
24. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Gtermany, 8–14 September 2018; pp. 3–19.
25. 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.
26. Codella, N.; Rotemberg, V.; Tschandl, P.; Celebi, M.E.; Dusza, S.; Gutman, D.; Helba, B.; Kalloo, A.; Liopyris, K.; Marchetti, M.; et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv* **2019**, arXiv:1902.03368.
27. Yuan, Y.; Lo, Y.C. Improving dermoscopic image segmentation with enhanced convolutional-deconvolutional networks. *IEEE J. Biomed. Health Inform.* **2017**, *23*, 519–526. [CrossRef] [PubMed]
28. 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] [PubMed]
29. Al-Masni, M.A.; Al-Antari, M.A.; Choi, M.T.; Han, S.M.; Kim, T.S. Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks. *Comput. Methods Programs Biomed.* **2018**, *162*, 221–231. [CrossRef] [PubMed]
30. Sarker, M.M.K.; Rashwan, H.A.; Akram, F.; Banu, S.F.; Saleh, A.; Singh, V.K.; Chowdhury, F.U.; Abdulwahab, S.; Romani, S.; Radeva, P.; et al. SLSDeep: Skin lesion segmentation based on dilated residual and pyramid pooling networks. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, 16–20 September 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 21–29.
31. Bi, L.; Kim, J.; Ahn, E.; Kumar, A.; Feng, D.; Fulham, M. Step-wise integration of deep class-specific learning for dermoscopic image segmentation. *Pattern Recognit.* **2019**, *85*, 78–89. [CrossRef]
32. Zhang, Y.; Liu, H.; Hu, Q. Transfuse: Fusing transformers and cnns for medical image segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 14–24.
33. Dai, D.; Dong, C.; Xu, S.; Yan, Q.; Li, Z.; Zhang, C.; Luo, N. Ms RED: A novel multi-scale residual encoding and decoding network for skin lesion segmentation. *Med. Image Anal.* **2022**, *75*, 102293. [CrossRef] [PubMed]
34. Wang, Y.; Wei, Y.; Qian, X.; Zhu, L.; Yang, Y. DONet: Dual objective networks for skin lesion segmentation. *arXiv* **2020**, arXiv:2008.08278.
35. Gu, R.; Wang, G.; Song, T.; Huang, R.; Aertsen, M.; Deprest, J.; Ourselin, S.; Vercauteren, T.; Zhang, S. CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. *IEEE Trans. Med. Imaging* **2020**, *40*, 699–711. [CrossRef]
36. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 10347–10357.
37. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
38. Tajbakhsh, N.; Gurudu, S.R.; Liang, J. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Trans. Med. Imaging* **2015**, *35*, 630–644. [CrossRef]
39. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; Halvorsen, P.; de Lange, T.; Johansen, D.; Johansen, H.D. Kvasir-seg: A segmented polyp dataset. In Proceedings of the MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, Republic of Korea, 5–8 January 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 451–462.
40. 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] [PubMed]
41. 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.

42. Fan, D.P.; Ji, G.P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; Shao, L. Pranet: Parallel reverse attention network for polyp segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, 4–8 October 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 263–273.
43. Liu, Q.; Han, Z.; Liu, Z.; Zhang, J. HMA-Net: A deep U-shaped network combined with HarDNet and multi-attention mechanism for medical image segmentation. *Med. Phys.* **2023**, *50*, 1635–1646. [CrossRef]
44. 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.

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.