Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Feature Ensemble Network for Medical Image Segmentation with Multi-scale Atrous Transformer

Di Gai[1] | **Yuhan Geng**[3] | **Zheng Huang**[1] | **Ruihua Zhou**[2] | **Xin Xiong**[4] | **Xia Huang**[4] | **Qi Wang**[1]

[1]School of Mathematics and Computer Sciences, Nanchang University, Nanchang, China

[2]School of Software, Nanchang University, Nanchang, China

[3]School of Public Health, University of Michigan, Ann Arbor, United States

[4]The First Affiliated Hospital, Jiangxi Medical College, Nanchang University, Nanchang, China

Correspondence

Corresponding author Zheng Huang, Nanchang University.

Email: huangzhengncu@outlook.com

Present address

This is sample for present address text this is sample for present address text.

Abstract

Recent years have witnessed notable advancements in medical image segmentation through deep convolutional neural networks. However, a notable limitation lies in the local operation of convolution, which hinders the ability to fully exploit global semantic information. To overcome the challenges prevalent in medical image segmentation, we propose the feature ensemble network with multi-scale atrous transformer. At the core of our approach lies the Multi-scale Contextual Integration module, which is based on the Multi-scale Atrous Transformer and facilitates contextual integration of multi-level features. To extract discriminative fine-grained features of the target region, we have incorporated a Hybrid Attention mechanism that synergistically combines spatial and channel attention, thereby sharpening the model's focus on crucial target information within high-level features. Additionally, we introduce the Channel-aware Feature Reconstruction module as an innovative component engineered to tackle feature similarity issues across different categories. This module performs feature reconstruction based on channel perception, effectively widening the feature gap between categories and enhancing the segmentation capability. It is worth mentioning that our approach surpasses the state-of-the-art method using three benchmark datasets in medical image segmentation.

KEYWORDS

Feature Ensemble, Medical Image Segmentation, Hybrid Attention, Atrous Transformer

1 | INTRODUCTION

In recent years, the field of medical imaging has witnessed a significant transformation due to advancements in digital technology and computational methods[1,2,3,4]. Among these developments, medical image segmentation has emerged as a critical component, offering a novel dimension in understanding complex anatomical and pathological structures[5,6,7,8,9]. Medical image segmentation involves the process of partitioning digital medical images into multiple segments to simplify and enhance the representation of an image into something more meaningful and easier to analyze. It is particularly pivotal in delineating regions of interest, such as tissues, organs, and pathological entities, from medical images like X-rays, MRI, CT scans, and ultrasound. The importance of this technology stems from its ability to provide detailed visual representations, which are essential for accurate diagnosis, planning and guidance of medical procedures, and monitoring treatment responses. For instance, in oncology, precise segmentation of a tumor from the surrounding tissues can significantly influence the course of treatment, ranging from radiotherapy planning to surgical interventions.

Conventional approaches for medical image segmentation encompass threshold-based methods[10], region-growth-based techniques[11], edge-detection-based algorithms[12], and others. However, these methods often lack generalizability as they require designing distinct features for different modalities of medical images. Additionally, their segmentation performance is limited when dealing with complex structures. With the rapid development of deep learning technology, we have witnessed the remarkable success of deep convolutional neural network based medical image segmentation methods[13,14,15,16]. Deep convolutional neural networks automatically learn feature representations through the use of convolutional kernels. Due to their higher adaptability, greater robustness, and superior performance compared to traditional methods, deep convolutional neural network-based approaches have been extensively

Abbreviations: ANA, anti-nuclear antibodies; APC, antigen-presenting cells; IRF, interferon regulatory factor.

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1049/ipr2.13157

This article is protected by copyright. All rights reserved.

Author Manuscript

applied in the field of medical image segmentation. However, convolutional operations in medical image segmentation have certain limitations. Firstly, the convolutional operation exhibits a preference for local information and struggles to establish connections between global features. Secondly, distinguishing between target and background in medical images poses a greater challenge, making it more difficult for models to focus on the target region and resulting in similar features across different categories.

To address the limitations of convolution in global modeling, numerous approaches have incorporated Vision Transformer into medical image segmentation tasks. Existing Vision Transformer-based networks[17,18,19,20,21,22] for medical image segmentation, such as TransUnet[23], MTUNet[24], which combine CNN to extract local features and employ Vision Transformer for global relationship modeling. However, these methods fail to consider the semantic gap between different encoders. The SwinUnet is a U-shaped segmentation network based on the Swin Transformer, which incorporates pure Transformer architecture. However, the utilization of pure Transformer encoder tends to overlook local feature information, resulting in diminished segmentation efficacy at the target's periphery. Consequently, addressing the acquisition of local information within global context emerges as a novel challenge.

To address the above challenges, this paper proposes the feature ensemble medical image segmentation model with multi-scale atrous transformer. We design a novel Multi-scale Contextual Integration (MSCI) module for integrating contextual multi-level features to deal with feature loss caused by semantic gaps between different encoders. The MSCI module is devised by Multi-scale Atrous Transformer (MSAT) block at different levels and performing self-attention computation, which can focus on local information during the global modelling. To enhance the interest on the target region of model, we propose a Hybrid Attention (HA) mechanism that integrates both spatial and channel attention mechanisms, enabling the model to gain deeper insights into the information about the target within deep features. In addition, we propose the channel-aware feature reconstruction (CFR) module that exploit channel awareness to improve feature reconstruction. This method can effectively amplify the differences between various categories of features and solve the problem of similarity between different categories, thus raising the final segmentation performance. We have conducted extensive experiments on three benchmark datasets for medical image segmentation, and the experimental results show that our method achieves state-of-the-art in both qualitative and quantitative analyses. The main contributions of this paper are summarized as follows:

(1) We propose a novel deep learning framework that integrates a feature aggregation network with a Multi-Scale Atrous Transformer (MSAT), thereby significantly enhancing the capture of global information and fine structural features in medical images.

(2) The Multi-Scale Context Integration (MSCI) module effectively merges local and global data, leading to improved identification of complex structures.

(3) To enable precise segmentation, we introduce a hybrid attention mechanism that combines spatial and channel attention to refine the model's focus on crucial target information.

(4) By incorporating a Channel-aware Feature Reconstruction module, our approach greatly enhances the model's ability to distinguish between different classes by improving feature separability and accurately reflecting inter-class differences.

2 | RELATED WORK

2.1 | Medical Image Segmentation

Medical image segmentation[25,26,27] is a high-profile research area in the field of medical image processing, which aims to automatically or semi-automatically segment critical structures and tissues to help doctors make more accurate diagnostic and therapeutic decisions. Earlier, Ronneberger et al.[15] proposed the U-Net architecture, which can efficiently capture feature information at different scales while having a small number of parameters, thus mitigating the overfitting problem. Zhou et al.[28] proposed UNet++, a variant of UNet, which alleviates the issue of unknown network depth by using an efficient collection of U-Nets of different depths that partially share an encoder, and designing a pruning scheme to speed up the inference of UNet++. Unlike traditional encoder-decoders, CE-Net[5] introduces a context extractor module for better capturing the contextual information in the image, which contributes to the processing of details and structural variations in medical images. Zhang et al.[29] suggested Edge-aTtention guidance Network, which embeds edge-attention representations to navigate the segmentation network, and the edge guidance module is used to learn edge attention representations at the early encoding layer, which are then transmitted to the multiscale decoding layer and integrated using a weighted aggregation module. Zhu et al.[30] introduced a novel segmentation technique that leverages both deep semantic understanding and edge information from multimodal MRI images, offering a comprehensive approach to accurately delineating brain tumors. Li et al.[31] proposed X-Net framework, which presents a dual encoding-decoding architecture that enhances medical image segmentation by effectively capturing and integrating complex spatial and contextual information across different scales. He et al.[32] proposed a cloud computing-support medical image segmentation method that utilizes multi-feature interaction and fusion, highlighting the scalability and computational efficiency of leveraging cloud

Author Manuscript

This article is protected by copyright. All rights reserved.

Author Manuscript

resources in medical imaging analysis. By focusing on multi-dimensional statistical features, Xu et al.[33] advanced medical image segmentation by providing a method that enriches the feature space for segmentation tasks, potentially improving accuracy and robustness in various applications.

2.2 | Transformer in Medical Image

Recently, researchers have endeavored to employ the Transformer architecture[34] for medical image segmentation tasks. By leveraging the self-attention mechanism[35] inherent in Transformers, this model adeptly captures intricate interrelationships among different regions within an image, thereby facilitating more precise organ segmentation. Notably, this approach exhibits superior capability in handling long-range dependencies compared to conventional CNN-based segmentation methods. TransUNet[23] is a fusion of Transformer and U-Net architectures for medical image segmentation, utilizing Transformers to capture global contexts by encoding tokenized image patches from a CNN feature map. Swin-Unet[36] utilized hierarchical Swin-Transformer in the encoder with shifted windows for context feature extraction, and a symmetric Swin Transformer-based decoder with a patch expanding layer for up-sampling, aiming to restore the spatial resolution of the feature maps. Wang et al.[24] introduced a novel Mixed Transformer Module that integrates Local-Global Gaussian-Weighted Self-Attention for efficient self-affinity calculation and External Attention for exploring interconnections, enabling the development of an MT-UNet model with a U-shaped architecture for accurate medical image segmentation through simultaneous learning of inter- and intra-affinities. Huang et al.[37] proposed ScaleFormer, a novel backbone architecture featuring two innovative designs. The first design incorporates a scale-wise intra-scale transformer that integrates CNN-based local features with transformer-based global cues, utilizing a lightweight Dual Axis MSA to extract row-wise and column-wise global dependencies within each scale. The second design introduces a spatial-aware inter-scale transformer that facilitates interaction among consensual regions in multiple scales, effectively highlighting cross-scale dependencies and addressing intricate scale variations. Wang et al.[38] employed a Channel Cross Transformer sub-module and a Channelwise Cross-Attention sub-module to conduct multi-scale channel fusion with Transformer and guide the fused channel-wise information to effectively connect with decoder features, addressing semantic gaps and enhancing accuracy in automatic medical image segmentation.

3 | METHOD

3.1 | Overview

The pipeline of our proposed netowrk is depicted in Fig. 1. The input image undergoes feature extraction through the encoder to get multi-level features, and the MSCI module effectively utilizes these multi-level features for feature integration, enhancing the discriminative capabilities of the decoder. The HA module combines the spatial attention mechanism and the channel attention mechanism to weight the underlying features extracted by the encoder in order to allow the decoder to make better segmentation decisions. The purpose of the CFR module is to perform feature separation and reconstruction to optimize the feature representation while removing feature redundancy. The loss function in our approach utilizes cross-entropy and Diceloss weighting, which are obtained by the following procedure:

$$\mathcal{L}_{total} = \eta_1 \mathcal{L}_{ce} + \eta_2 \mathcal{L}_{dice}, \tag{1}$$

where $\mathcal{L}_{ce}$ and $\mathcal{L}_{dice}$ represent cross-entropy loss and Diceloss, respectively, and η_1 and η_2 are the weights of cross-entropy loss and Diceloss, respectively.

3.2 | Multi-scale Context Integration Module

In the encoder-decoder architecture, the encoder is capable of extracting multi-level features that encompass both low-level edge information and high-level semantic information. The proposed method effectively captures diverse features of the target at different scales, enabling better adaptation to varying sizes of lesion areas, structures, and organs. The MSCI module can effectively utilize the four layers of features extracted from the backbone network for feature integration, which provides a more discriminative feature representation for the decoder component. The MSCI module mainly consists of four levels of Multi-scale Atrous Transformer (MSAT) blocks of backbone features, which process multi-level features sequentially from the deepest to the shallowest level by oversizing the receptive field and capturing more contextual information to enhance the feature representation. More specifically, we employ the MSAT module to process and integrate the highest level backbone features, which are the last convolution-extracted features, with the previous level backbone features as input for the subsequent MSAT Block:

$$F_i = MSAT_i(P_i, F_{i-1}), \tag{2}$$

where P_i is the ith layer of backbone features and F_{i-1} is the processing result of the $i-1th$ MSAT block.

This article is protected by copyright. All rights reserved.

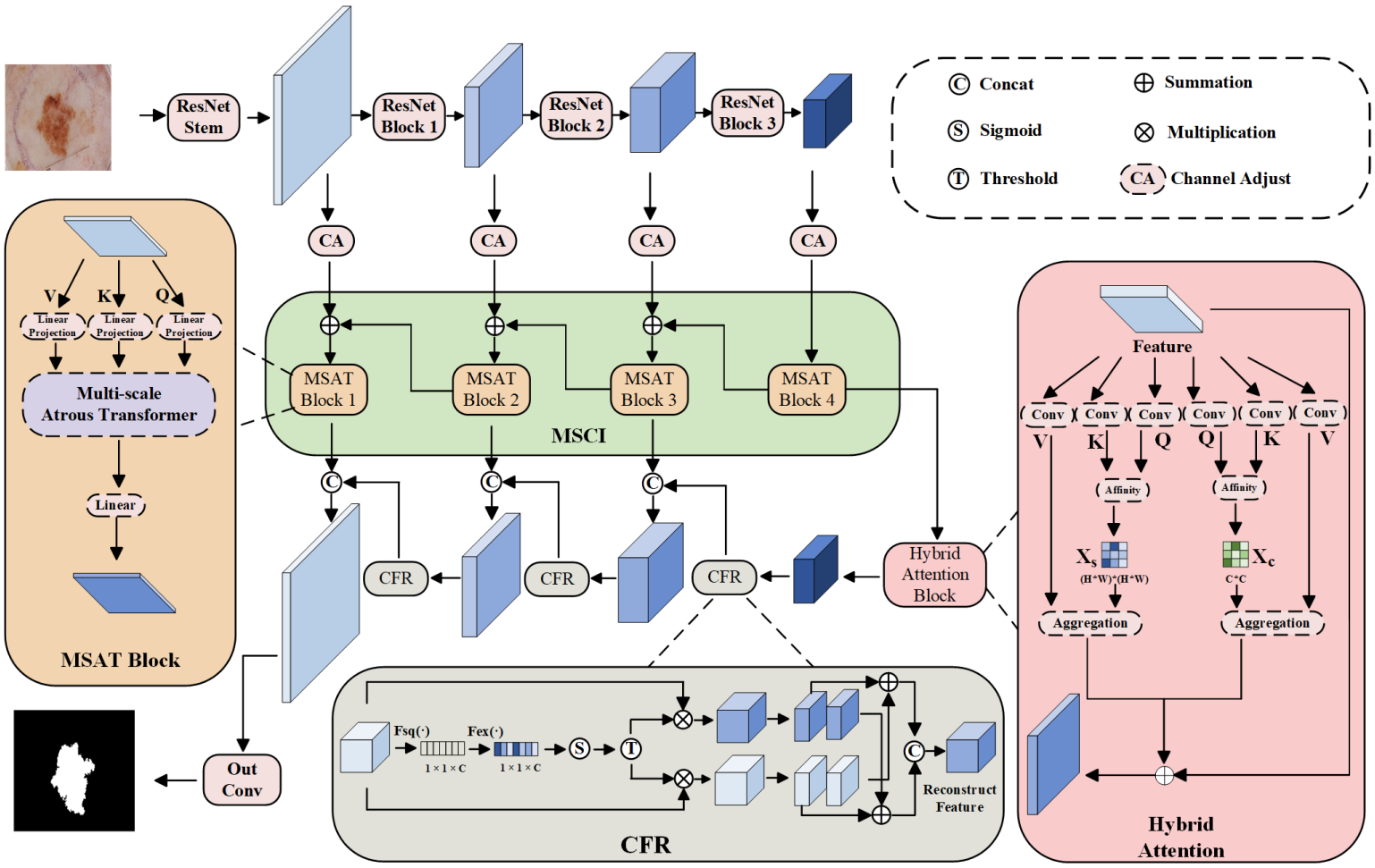


FIGURE 1 Overview of our proposed network.

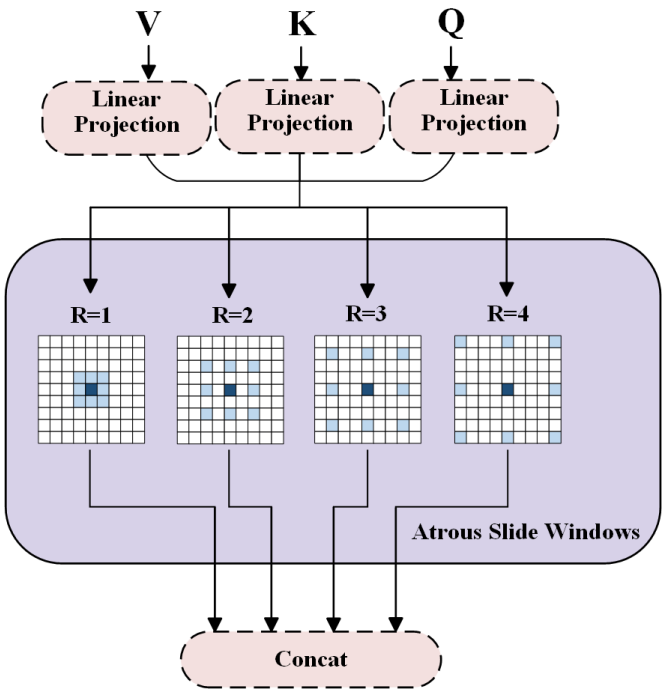


FIGURE 2 The details of the atrous slide windows.

Multi-scale Atrous Transformer. Numerous Transformer-based models for medical image segmentation, such as TransUnet[23], incorporate a traditional Transformer model to segment the image into multiple blocks and perform self-attention computation. However, this approach has limitations in adapting to targets of varying sizes. In contrast, SwinUnet[36] utilizes a multi-scale sliding window mechanism to achieve self-attention, enabling better fusion of information from different scales and establishing connections between regions through sliding windows. Differing from these approaches, we propose a novel atrous Transformer that demonstrates improved performance in medical image segmentation tasks. A detailed presentation of the MSAT block is shown in Fig. 2. The Query, Key, and Value matrices are initially obtained through three different linear projections for the input feature. Subsequently, the feature map's channels are divided into n distinct heads, enabling the execution of self-attention mechanism within each head with varying sampling radii. Specifically, different scales are employed by Key and Value to sparsely select patches, followed by performing self-attention on these selected patches, which are obtained by the following procedure:

$$X = ASW(Q, K, V, r), \tag{3}$$

where r is the radius of sparse sampling. For the self-attention mechanism of each ASW, we formalize this process as:

Author Manuscript

This article is protected by copyright. All rights reserved.

$$x_{ij} = Softmax(\frac{q_{ij}K_r^T}{\sqrt{d_k}})V_r, 1 \leq i \leq W, 1 \leq j \leq H, \quad (4)$$

where d_k represents the feature dimension. For the query at position (i, j) in the feature map, ASW sparsely selects Key and Value to sample in a sliding window of size $w \times w$ centered on (i, j) and performs self-attention calculation. We reformulate Equation (5) and Equation (6) as follows:

$$t_i = ASW_i(Q_i, K_i, V_i, r_i), 1 \leq i \leq n, \quad (5)$$

$$X = Linear(Concat[t_1, ..., t_n]), \quad (6)$$

where t_i is the self-attention result under the $i-th$ atrous window, Q_i, K_i and V_i represent Query, Key and Value respectively, r_i is the radius of sampling, and a total of n groups are used. By setting different sampling radii for different heads, *ASW* explicitly satisfies the locality and sparsity properties, which can effectively model long-range dependencies.

3.3 | Hybrid Attention Module

In order to strengthen the role of important features, we propose a novel HA mechanism with two branches, spatial attention and channel attention, which optimizes the feature representation both spatially and channel-wise. As shown in Fig. 1, our HA module has two branches, the spatial branch and the channel branch. The final result is obtained by merging the spatial and channel branches.

Channel Branch. The input feature F, belonging to $\mathbb{R}^{C\times H\times W}$ in the channel branch, is first reshaped to obtain query, key, and value. Here, q, k, v belongs to $\mathbb{R}^{C\times N}$ with $N = H \times W$. Subsequently, matrix multiplication is performed between Q and K transpose followed by applying softmax for calculating channel attention. This results in obtaining the attention map as shown in Equation (7):

$$x_{ij} = \frac{Q_i \cdot K_j}{\sum_{j=1}^{C} exp(Q_i \cdot K_j)}, \quad (7)$$

In order to improve the adaptive ability, we add a learnable parameter λ so that we get the result of channel attention branching as shown in Equation (8):

$$F_c = \lambda \sum_{j=1}^{C}(x_{ij}V_j), \quad (8)$$

where F_c denotes the result obtained by the feature after channel branching and C is the number of channels of the feature.

Spatial Branch. The features in the spatial branch are processed using a 1x1 convolution to obtain q, k, and v. Here, $\{q, k\}$ belongs to $\mathbb{R}^{C_1\times N}$, where $C_1 = C/8$, and V belongs to $\mathbb{R}^{C\times N}$. This procedure can be formulated as follows:

$$x_{ij} = \frac{Q_i \cdot K_j}{\sum_{j=1}^{N} exp(Q_i \cdot K_j)}, \quad (9)$$

where Q_i denotes the $i-th$ column of matrix Q and x_{ij} denotes the effect of the $j-th$ feature on the $i-th$ feature.

Subsequently, matrix multiplication is performed between our Q and K transpositions followed by applying softmax to generate the spatial attention map $X' = \mathbb{R}^{N\times N}$. The transpositions of V and X' are subjected to a matrix multiplication, resulting in the transformation to $\mathbb{R}^{C\times H\times W}$. Similarly to the channel attention branch, we introduce a scaling parameter γ into the spatial attention branch for obtaining the final outcome as depicted in Equation (10):

$$F_s = \gamma \sum_{j=1}^{N}(x_{ij}V_j), \quad (10)$$

where F_s denotes the result obtained by the features after spatial branching and N is the number of features.

The features are obtained by aggregating the outcomes of spatial branching and the attention mechanism in channel branching, yielding an expression as follows:

$$F' = F + F_s + F_c \quad (11)$$

where F' denotes the result obtained by feature F after HA.

3.4 | Channel-aware Feature Reconstruction

The CFR module employs channel perception for feature reconstruction, while the objective of channel-aware separation is to segregate information-rich feature maps from less informative ones using channel perception. Specifically, given a feature F belonging to $C \times H \times W$, we obtain the feature map F' of $C \times 1 \times 1$ through convolution operation followed by applying the sigmoid function to derive the weights for each channel. We then filter out channels with larger and smaller weights based on a threshold T, as illustrated in Equation (12):

$$f' = Sigmoid(F_{ex}(F_{sq}(f))), \quad (12)$$

where $F_{sq}(-)$ adjusts the feature f to $1 \times 1 \times C$, $F_{ex}(-)$ evaluates the significance of each channel based on the outcomes of $F_{sq}(-)$, and subsequently obtains the weight for each channel using Sigmoid.

We multiply the original features and mask to obtain the information-rich feature $T_1(f')$ and the less informative feature $T_2(f')$ as shown in follows:

$$T_1(f') = \begin{cases} 0 & \text{if } f_i' < V \\ 1 & \text{if } f_i' \geq V \end{cases} 1 \leq i \leq C, \quad (13)$$

Author Manuscript

This article is protected by copyright. All rights reserved.

$$T_2(f') = \begin{cases} 1 & \text{if } f_i' < V \\ 0 & \text{if } f_i' \geq V \end{cases} 1 \leq i \leq C. \tag{14}$$

After multiplying with the original features, we separated two features as shown in follows:

$$W_1 = T_1(f') \cdot f, \tag{15}$$

$$W_2 = T_2(f') \cdot f, \tag{16}$$

where W_1 is the feature with rich information and W_2 is the feature with less information. To enhance the overall information richness, we perform cross-organization on these two features. As illustrated in Fig. 1, we divide W_1 and W_2 equally along the channel dimension and subsequently concatenate them. Finally, a convolution operation is applied to obtain the resultant feature map according to follows:

$$\begin{aligned} W_1' &= W_1^{0,C/2} + W_2^{C/2,C} \\ W_2' &= W_2^{0,C/2} + W_1^{C/2,C}, \end{aligned} \tag{17}$$

$$f_w = Conv(Concat(W_1', W_2')), \tag{18}$$

where W_1' and W_2' are the refined features, f_w is the resultant feature map.

4 | EXPERIMENTS

4.1 | Implemention Details and Data Set

Implemention Details. The experiments are conducted exclusively using PyTorch in an experimental environment running Ubuntu 18.04. A total of 2,000 epochs are trained on a Tesla V100 32GB GPU with a batch size of 8. All images are resized to a resolution of 256 $\times$ 256 pixels. The Adam optimizer is employed for optimizing the network parameters, initializing the learning rate at 0.0001. The hyperparameters λ_1 and λ_2 used to balance the loss are set to 0.5 and 0.5, respectively. To quantitatively evaluate the model's performance, we adopt commonly used evaluation metrics including Jaccard Index (JA), Dice coefficient (DI), pixelwise Accuracy (AC), Sensitivity (SE), Specificity (SP), and Hasdiff 95 (HD95).

ISIC-2017 dataset. The ISIC dataset contains a total of 5423 images of skin lesion areas with different scale sizes, different shapes, and different colors, of which 3461 images are selected as the training set and 2002 images serve as the test set.

Kvasir Dataset. The gastrointestinal polyp segmentation dataset is obtained from Kvasir-SEG, comprising of 1000 annotated images with varying dimensions ranging from 332×487 to 1920×1072 pixels. To ensure unbiased evaluation, we perform a random split of the dataset into a training set consisting of 700 images and a test set containing the remaining 300 images.

LGG dataset. The LGG dataset comprises 110 MR image slices of low-grade gliomas, sourced from the Cancer Genome Atlas and the Cancer Imaging Archive. Following dataset processing, a total of 1311 images are collected for experimentation purposes, with 1049 images randomly allocated for training and 262 for testing.

4.2 | Comprasion Experiments

To validate the efficacy of our proposed approach, we perform a comparative analysis with several state-of-the-art medical image segmentation methods, including UNet[15], AttentionUNet[39], UNet++[28], MTUNet[24], MedT[40], ScaleFormer[37], SwinUnet[36], TransUnet[23], UCTransNet[38], Transception[41].

We assess the performance of the proposed segmentation network using the ISIC dataset and compare it with other existing methods. To enhance visualization and facilitate presentation, we carefully select three skin pigmented tumor samples, showcasing their corresponding segmentation results in Figure 3. In Figure 4, the segmentation results of sample (a) demonstrate that our method successfully identifies the target amidst interference, while UNet, UNet++, TransUnet, SwinUnet, and Transception fail to accurately detect the pigmented tumor region. This is because compare with TransUnet and Transception, our method can learn feature representations more carefully by computing self-attention through multi-scale sampling. For sample (b), our method precisely captures the target's edge due to our MSCI module's ability to optimize feature representation by integrating contextual multi-level features. In case of irregularly shaped sample (c), our method achieves more accurate segmentation results owing to effective feature reconstruction facilitated by our CFR module, enabling better foreground-background discrimination. Furthermore, we quantitatively analyze these outcomes in Table 1 and observe that compared to UCTransNet, our method outperforms others in all evaluation metrics: JA metric improves by 1.31%, DI metric improves by 0.96%, AC metric improves by 0.41%, SE improves by 1.52%, SP improves by 1.28%, and HD95 is enhanced by 0.26/pixel.

To validate the applicability of our model across diverse medical image datasets, we assessed the proposed segmentation network on the Kvasir-SEG dataset and compared it with alternative methods. Fig. 4 illustrates three samples of intestinal polyps along with their corresponding segmentation results. In Fig. 4, for sample (a), our method demonstrates superior ability to segment the target region comprehensively, whereas UNet, SwinUnet, and UCTransNet fail to accurately identify the polyp region due to our HA Mechanism Module's enhanced focus on target features. For sample (b), our model achieves segmentation results closest to ground truth labels compared to Medical Transformer, MTUNet, and ScaleFormer which fall short in

Author Manuscript

This article is protected by copyright. All rights reserved.

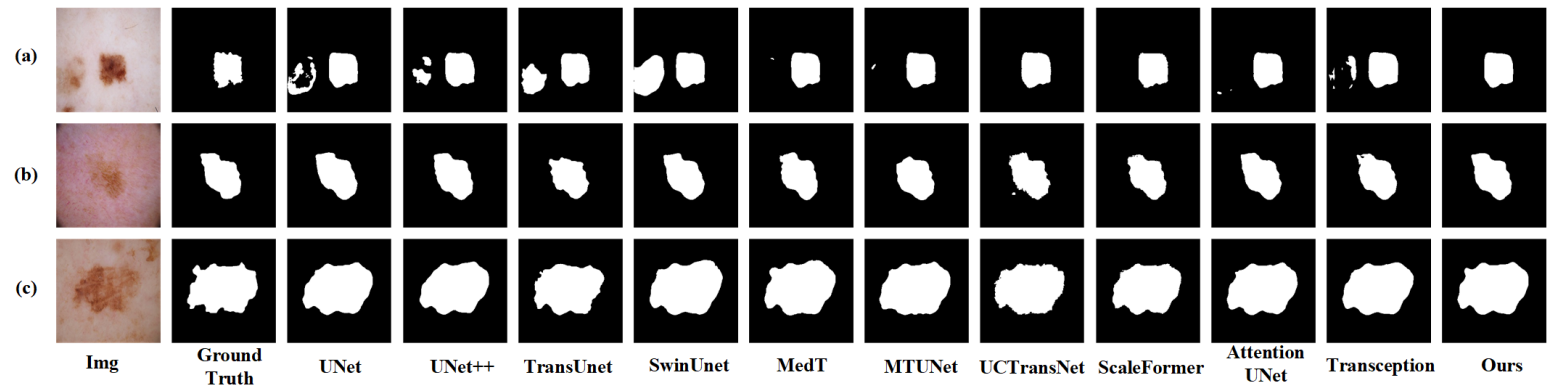


FIGURE 3 The visualization of different methods on the ISIC dataset.

TABLE 1 Performance of different segmentation methods on the test set of the ISIC dataset.

Methods	JA(%)↑	DI(%)↑	AC(%)↑	SE(%)↑	SP(%)↑	HD95(/pixel)↓
UNet [15]	86.75	92.21	96.14	96.22	97.48	1.89
AttentionUNet [39]	85.48	91.27	95.72	95.14	97.84	2.17
UNet++ [28]	86.84	92.24	96.26	96.40	97.98	1.85
MTUNet [24]	84.87	90.85	95.60	96.01	98.24	2.20
MedT [40]	82.28	89.24	94.90	93.67	97.56	2.49
ScaleFormer [37]	86.05	91.82	95.93	96.41	97.84	2.05
SwinUnet [36]	85.03	91.06	95.69	96.30	97.96	1.87
TransUnet [23]	87.01	92.30	96.33	97.52	97.50	1.94
UCTransNet [38]	87.33	92.56	96.39	96.65	97.56	1.78
Transception [41]	87.10	92.47	96.29	97.81	97.90	1.80
Ours	**88.64**	**93.52**	**96.80**	**98.17**	**98.84**	**1.52**

TABLE 2 Performance of different segmentation methods on the test set of the Kvasir dataset.

Methods	JA(%)↑	DI(%)↑	AC(%)↑	SE(%)↑	SP(%)↑	HD95(/pixel)↓
UNet [15]	79.82	87.41	95.20	85.58	98.21	3.96
AttentionUNet [39]	80.59	87.34	95.48	86.23	98.42	4.09
UNet++ [28]	80.97	86.99	95.11	85.71	98.07	4.68
MTUNet [24]	79.76	87.93	94.85	86.75	96.75	5.48
MedT [40]	80.55	87.91	95.29	85.36	97.82	5.99
ScaleFormer [37]	79.50	87.87	94.08	83.38	96.95	6.19
SwinUnet [36]	80.73	86.31	95.34	85.01	98.16	4.99
TransUnet [23]	80.76	86.90	95.77	86.22	98.99	4.24
UCTransNet [38]	80.56	86.46	95.43	87.14	98.49	3.12
Transception [41]	81.59	88.15	95.99	88.57	97.59	5.88
Ours	**83.66**	s**89.45**	**96.75**	**90.70**	**99.03**	**2.92**

complete target delineation. Even when dealing with sample (c) where the target closely resembles background features, our approach still yields improved segmentation outcomes owing to feature filtering and reconstruction facilitated by our CFR module that enhances class discrimination within features. Additionally, we quantitatively analyze the performance of various methods on the Kvasir-SEG dataset as presented in Table 2. The data from Table 2 reveals that across all evaluation metrics, our model outperforms other approaches; specifically compared to Transception: JA metric improves by 2.07%, DI metric improves by 1.30%, AC metric improves by 1.09%, SE improves by 2.13%, SP improves by 1.44%, and HD95 improves by 0.33/pixel respectively. The aforementioned visual analysis as well as quantitative assessment substantiate that our method attains state-of-the-art performance.

To evaluate the performance of our method on MRI datasets, we conducted an assessment of the proposed segmentation network using the LGG dataset and compared it with other methods. Additionally, we visually analyzed three brain tumor MRI images and presented the segmentation results in Figure 5. In sample (a), UNet++ and MTUNet failed to completely segment the interior of the tumor, while our model effectively captures multi-scale contextual features through the MSCI module, resulting in more accurate target segmentation by capturing finer image details. For irregularly shaped samples (b) and (c), our model also outperforms other models in terms of detailed segmentation. The visualization analysis confirms that our model achieves superior results on MRI datasets. Furthermore, we quantitatively evaluated the segmentation results on the LGG dataset as shown in Table 3. By observing data of the table, it is evident that our method surpasses existing state-of-the-art methods across all evaluation metrics. Compared to UCTransNet, our method demonstrates a 0.28% improvement in JA metric, a 0.14% improvement in DI metric, and a 0.11/pixel improvement in HD95. This is because UCTransNet does not consider multi-scale self-attention calculation, while our method achieves better segmentation effect through multi-scale atrous self-attention calculation.

4.3 | Ablation Study

Ablation of individual modules. To evaluate the impact of our proposed modules on model performance, we conducted ablation experiments where specific modules were removed to observe changes in performance. The results are presented in Table 4. In Scheme 1, when the CFR module was deleted while keeping all other factors constant, there was a decrease of 0.8% in the JA metric for quantized segmentation results on the Kvasir-SEG dataset. In Scheme 2, removing the HA

Author Manuscript

This article is protected by copyright. All rights reserved.

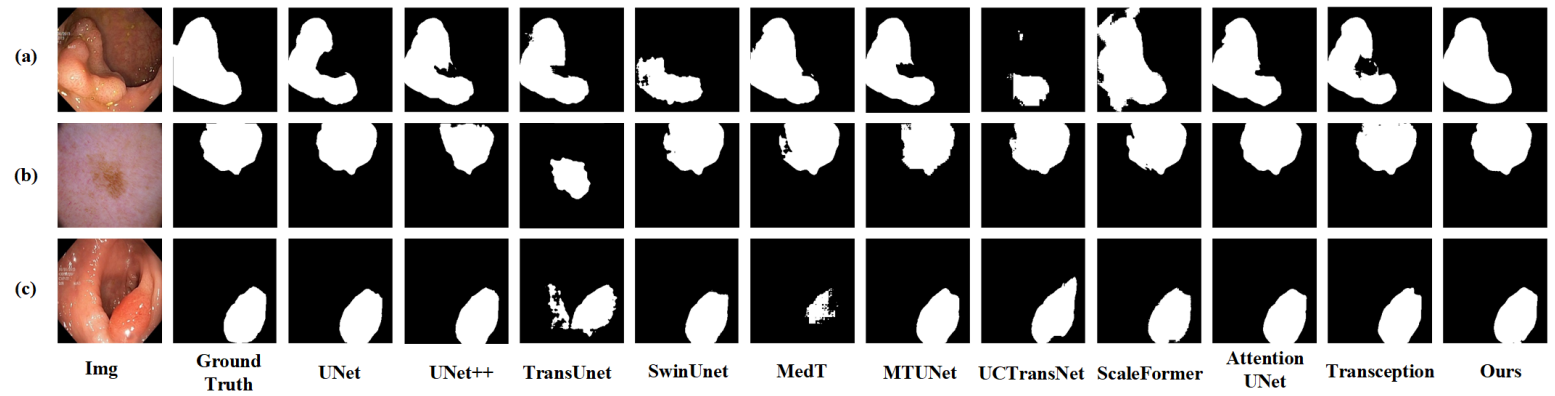


FIGURE 4 The visualization of different methods on the Kvasir dataset.

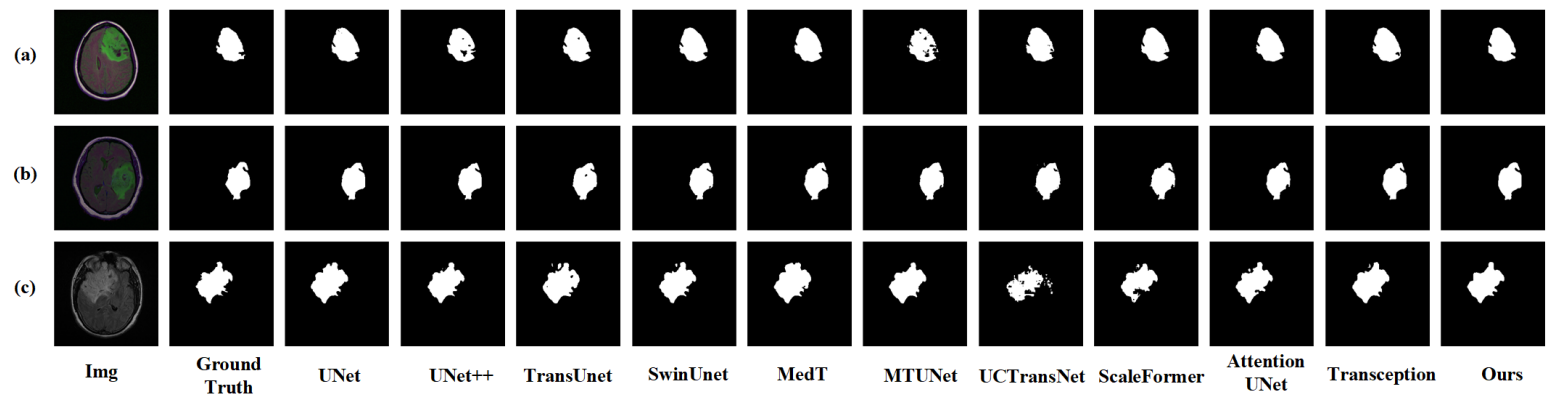


FIGURE 5 The visualization of different methods on the LGG dataset.

TABLE 3 Performance of different segmentation methods on the test set of the LGG dataset.

Methods	JA(%)↑	DI(%)↑	AC(%)↑	SE(%)↑	SP(%)↑	HD95(/pixel)↓
UNet [15]	86.57	91.81	99.21	93.07	99.65	2.38
AttentionUNet [39]	87.93	93.43	99.36	91.89	99.77	2.91
UNet++ [28]	88.16	93.54	99.36	95.47	99.84	1.15
MTUNet [24]	87.92	93.41	99.34	94.57	99.84	1.32
MedT [40]	80.84	88.83	98.82	86.87	99.83	2.55
ScaleFormer [37]	87.42	93.15	99.33	95.44	99.89	1.19
SwinUnet [36]	82.54	90.11	99.03	89.99	99.76	2.13
TransUnet [23]	87.30	93.03	99.31	94.50	99.83	1.26
UCTransNet [38]	88.35	93.68	99.38	95.27	99.81	1.19
Transception [41]	86.42	92.48	99.25	93.93	99.89	1.30
Ours	**88.63**	**93.82**	**99.40**	**95.57**	**99.92**	**1.08**

module resulted in a decrease of 0.61% for the AC metric and 1.04% for the SP metric. Scheme 3 involved removing the MSCI module, which significantly affected model performance with a decrease of 2.2% in JA metric and a reduction of HD95 metric by 1.72/pixel. Finally, when all modules were removed to obtain baseline performance, it became evident that each proposed module effectively contributed to improving segmentation accuracy.

To investigate the impact of sample group and distance in the Atrous Transformer within our model, we conducted ablation experiments, and the results are presented in Table 5. It is evident that a single sample leads to lower model accuracy, with a decrease of 1.61% in JA index and 1.28% in DI index compared to using four samples. When increasing the sample number to two, there was some improvement in performance; however, different sampling radii had varying effects on the results. Notably, models utilizing sampling radii of 2 and 3 outperformed those with radii of 1 and 2 or those with radii of 3 and 4. Furthermore, employing three samples yielded more accurate results than using two samples alone. The best performance was achieved when utilizing four samples; compared to models with three samples and a sampling radius set as (1,2,3), JA improved by an additional 0.45%, DI improved by an additional 0.42%, while HD95 also showed a notable improvement of approximately +0.35/pixel.

Finally, a series of ablation experiments were conducted to validate the role of the spatial branch and channel branch in our proposed HA module. The experimental results are presented in Table 6. In the first experiment, we removed the spatial branch, resulting in a decrease of 0.42% in the JA indicator, 0.34% in the DI indicator, and 0.61% in the AC indicator according to the data shown in Table 6. These findings indicate that incorporating the spatial branch contributes to improved model accuracy. In the second experiment, we eliminated the channel branch which led to a reduction of 0.56% in JA indicator, 0.52% in DI indicator, and 0.54% in AC indicator as indicated by table data analysis. The aforementioned experiments demonstrate

Author Manuscript

This article is protected by copyright. All rights reserved.

TABLE 4 Ablation study of each modules.

Methods	MSCI	HA	CFR	JA(%)↑	DI(%)↑	AC(%)↑	SE(%)↑	SP(%)↑	HD95(/pixel)↓
Baseline				80.67	86.82	95.30	84.87	98.06	5.33
Scheme.1	✓	✓		82.86	88.76	96.34	88.80	98.48	3.40
Scheme.2	✓		✓	82.97	88.89	96.14	89.04	97.99	3.82
Scheme.3		✓	✓	81.49	87.42	95.74	86.83	98.07	4.64
Ours	✓	✓	✓	**83.66**	**89.45**	**96.75**	**90.70**	**99.03**	**2.92**

TABLE 5 Ablation study of different sampling group and distance.

Group	Distance	JA(%)↑	DI(%)↑	AC(%)↑	SE(%)↑	SP(%)↑	HD95(/pixel)↓
1	(1)	82.05	88.17	95.86	87.59	97.99	4.17
2	(1,2)	82.11	88.21	96.16	87.76	98.18	3.86
2	(2,3)	82.37	88.43	96.12	87.93	98.21	3.62
2	(3,4)	82.24	88.35	96.01	88.04	98.36	3.79
3	(1,2,3)	83.21	89.03	96.21	88.45	98.42	3.27
3	(2,3,4)	82.84	88.92	96.37	88.16	98.73	3.41
4	(1,2,3,4)	**83.66**	**89.45**	**96.75**	**90.70**	**99.03**	**2.92**

that both spatial and channel branches have significant impact on enhancing model accuracy.

TABLE 6 Ablation study of spatial branch and channel branch.

Methods	JA(%)↑	DI(%)↑	AC(%)↑	SE(%)↑	SP(%)↑	HD95(/pixel)↓
w/o Spatial Branch	83.24	89.11	96.14	88.21	98.19	3.54
w/o Channel Branch	83.10	88.93	96.21	88.02	98.28	3.19
Ours	**83.66**	**89.45**	**96.75**	**90.70**	**99.03**	**2.92**

5 | CONCLUSIONS

In this paper, we propose the feature ensemble network for medical image segmentation. Notably, the MSCI module effectively addresses the semantic gap between CNN and Transformer encoders by integrating multi-scale contextual features through multi-scale cavity sampling Transformers, thereby considering the relationship between local features during global feature modeling. To enhance focus on target features, we introduce the HA module that facilitates spatial and channel-level attention. Additionally, our CFR module performs feature reconstruction based on channel perception to filter critical features more effectively, resulting in larger gaps between different feature categories and improved model discriminability. In future work, we aim to extend our model to 3D for volume segmentation.

AUTHOR CONTRIBUTIONS

Di Gai: Funding acquisition, Writing - original draft, Project administration. Yuhan Geng: Validation, Visualization. Zheng Huang: Methodology, Conceptualization. Ruihua Zhou: Writing - Original Draft, Formal analysis. Xin Xiong: Investigation, Supervision. Xia Huang: Supervision, Funding acquisition. Qi Wang: Supervision, Resources.

ACKNOWLEDGMENTS

This work was supported by the the Jiangxi Provincial Natural Science Foundation under Grant No. 20224BAB212011, 20232BAB212008 and 20232BAB202051, the Jiangxi Science and Technology Program under Grant No. 20232ABC03A23, 20232ABC03A32, 20232ABC03W01, the National Natural Science Foundation of China under Grant No. 82360341, and Science and Technology Program of National Health Commission under Grant No. NIHA23JXH023.

FINANCIAL DISCLOSURE

There are no financial conflicts of interest to disclose.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

References

1. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015:3431-3440

2. Gai D, Zhang J, Xiao Y, et al. GL-Segnet: Global-Local representation learning net for medical image segmentation. *Frontiers in Neuroscience.* 2023;17:1153356. doi: 10.3389/fnins.2023.1153356

3. Zhi L, Jiang W, Zhang S, Zhou T. Deep neural network pulmonary nodule segmentation methods for CT images: Literature review and experimental comparisons. *Computers in Biology and Medicine.* 2023;164:107321. doi: 10.1016/j.compbiomed.2023.107321

Author Manuscript

This article is protected by copyright. All rights reserved.

Author Manuscript

4. Hao D, Li H. A graph-based edge attention gate medical image segmentation method. *IET Image Processing.* 2023;17(7):2142-2157. doi: https://doi.org/10.1049/ipr2.12780

5. Gu Z, Cheng J, Fu H, et al. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE transactions on medical imaging.* 2019;38(10):2281-2292. doi: 10.1109/TMI.2019.2903562

6. Wang Z, Zhu J, Fu S, Mao S, Ye Y. RFPNet: Reorganizing feature pyramid networks for medical image segmentation. 2023;163:107108. doi: 10.1016/j.compbiomed.2023.107108

7. Valanarasu JMJ, Patel VM. Unext: Mlp-based rapid medical image segmentation network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. 2022:23-33

8. Xie X, Zhang W, Pan X, et al. Canet: Context aware network with dual-stream pyramid for medical image segmentation. *Biomedical Signal Processing and Control.* 2023;81:104437. doi: 10.1016/j.bspc.2022.104437

9. Gai D, Feng R, Min W, et al. Spatiotemporal Learning Transformer for Video-Based Human Pose Estimation. *IEEE Transactions on Circuits and Systems for Video Technology.* 2023;33(9):4564-4576. doi: 10.1109/TCSVT.2023.3269666

10. Khairuzzaman AKM, Chaudhury S. Multilevel thresholding using grey wolf optimizer for image segmentation. *Expert Systems with Applications.* 2017;86:64-76.

11. Adams R, Bischof L. Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 1994;16(6):641-647.

12. Wang L, Chen G, Shi D, et al. Active contours driven by edge entropy fitting energy for image segmentation. *Signal Processing.* 2018;149:27-35.

13. 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. 2016:770-778.

14. Li X, Chen H, Qi X, Dou Q, Fu CW, Heng PA. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Transactions on Medical Imaging.* 2018;37(12):2663-2674.

15. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. 2015:234-241

16. Wang S, Liu Y, Sun Y, Yin B. SACNet: Shuffling atrous convolutional U-Net for medical image segmentation. *IET Image Processing.* 2023;17(4):1236-1252. doi: https://doi.org/10.1049/ipr2.12709

17. Wang W, Xie E, Li X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021:568-578

18. Chen CFR, Fan Q, Panda R. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021:347-356

19. Han K, Xiao A, Wu E, Guo J, XU C, Wang Y. Transformer in transformer. In: . 34 of *Advances in Neural Information Processing Systems*. 2021:15908-15919

20. Zheng S, Lu J, Zhao H, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021:6881-6890

21. Xu W, Xu Y, Chang T, Tu Z. Co-Scale Conv-Attentional Image Transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021:9961-9970

22. Gai D, Zhang J, Xiao Y, Min W, Zhong Y, Zhong Y. RMTF-Net: Residual Mix Transformer Fusion Net for 2D Brain Tumor Segmentation. *Brain Sciences.* 2022;12(9). doi: 10.3390/brainsci12091145

23. Chen J, Lu Y, Yu Q, et al. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306.* 2021. doi: 10.48550/arXiv.2102.04306

24. Wang H, Xie S, Lin L, et al. Mixed Transformer U-Net for Medical Image Segmentation. In: IEEE International Conference on Acoustics, Speech and Signal Processing. 2022:2390-2394

25. Jin Q, Meng Z, Sun C, Cui H, Su R. RA-UNet: A hybrid deep attention-aware network to extract liver and tumor in CT scans. *Frontiers in Bioengineering and Biotechnology.* 2020;8:605132. doi: 10.3389/fbioe.2020.605132

26. Yuan F, Zhang Z, Fang Z. An effective CNN and Transformer complementary network for medical image segmentation. *Pattern Recognition.* 2023;136:109228. doi: 10.1016/j.patcog.2022.109228

27. Huang X, Deng Z, Li D, Yuan X, Fu Y. Missformer: An effective transformer for 2d medical image segmentation. *IEEE Transactions on Medical Imaging.* 2022;42(5):1484-1494. doi: 10.1109/TMI.2022.3230943

28. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Transactions on Medical Imaging.* 2020;39(6):1856-1867. doi: 10.1109/TMI.2019.2959609

29. Zhang Z, Fu H, Dai H, Shen J, Pang Y, Shao L. ET-Net: A Generic Edge-aTtention Guidance Network for Medical Image Segmentation. In: Proceeding of the Medical Image Computing and Computer Assisted Intervention. 2019:442–450.

30. Zhu Z, He X, Qi G, Li Y, Cong B, Liu Y. Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI. *Information Fusion.* 2023;91:376-387. doi: https://doi.org/10.1016/j.inffus.2022.10.022

31. Li Y, Wang Z, Yin L, Zhu Z, Qi G, Liu Y. X-Net: a dual encoding–decoding method in medical image segmentation. *The Visual Computer.* 2023;39. doi: 10.1007/s00371-021-02328-7

32. He X, Qi G, Zhu Z, Li Y, Cong B, Bai L. Medical image segmentation method based on multi-feature interaction and fusion over cloud computing. *Simulation Modelling Practice and Theory.* 2023;126:102769. doi: https://doi.org/10.1016/j.simpat.2023.102769

33. Xu Y, He X, Xu G, et al. A medical image segmentation method based on multi-dimensional statistical features. *Frontiers in Neuroscience.* 2022;16. doi: 10.3389/fnins.2022.1009581

34. Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need. In: . 30 of *Advances in Neural Information Processing Systems*. 2017

35. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. 2020

36. Cao H, Wang Y, Chen J, et al. Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation. In: European conference on computer vision. 2022:205-218

37. Huang H, Xie S, Lin L, et al. ScaleFormer: Revisiting the Transformer-based Backbones from a Scale-wise Perspective for Medical Image Segmentation. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22. 2022:964-971

This article is protected by copyright. All rights reserved.

38. Wang H, Cao P, Wang J, Zaiane OR. Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In: Proceedings of the AAAI conference on artificial intelligence. 2022:2441-2449

39. Oktay O, Schlemper J, Folgoc LL, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*. 2018. doi: 10.48550/arXiv.1804.03999

40. Valanarasu JMJ, Oza P, Hacihaliloglu I, Patel VM. Medical transformer: Gated axial-attention for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. 2021:36-46

41. Azad R, Jia Y, Aghdam EK, Cohen-Adad J, Merhof D. Enhancing Medical Image Segmentation with TransCeption: A Multi-Scale Feature Fusion Approach. *arXiv preprint arXiv:2301.10847*. 2023.

Author Manuscript

This article is protected by copyright. All rights reserved.