IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 42, No. 1, February, 2025, pp. 531-540
Journal homepage: http://iieta.org/journals/ts

MSF-TransUNet: A Multi-Scale Feature Fusion Transformer-Based U-Net for Medical Image Segmentation with Uniform Attention

Ying Jiang[1], Lejun Gong [1], Hao Huang[2], Mingming Qi[3*]

[1] School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
[2] State Key Laboratory of Digital Medical Engineering, School of Biological Science and Medical Engineering, Southeast University, Nanjing 210023, China
[3] School of Data Science and Artificial Intelligence, Wenzhou University of Technology, Wenzhou 325000, China

Corresponding Author Email: qmm19742021@163.com

https://doi.org/10.18280/ts.420145

Received: 27 July 2024
Revised: 8 November 2024
Accepted: 27 November 2024
Available online:28 February 2025

Keywords:
attention mechanisms, deep learning, medical imaging segmentation, multi-scale feature fusion

ABSTRACT

Accurate medical image segmentation is essential for computer-assisted diagnosis and treatment systems. While conventional U-Net architectures and hybrid models integrating U-Net with Transformer networks have demonstrated remarkable performance in automatic segmentation tasks, these approaches frequently face challenges in effectively integrating multi-scale features. Additionally, semantic inconsistencies arising from simple skip connections during the encoding-decoding process remain problematic. To address these limitations, a novel architecture, MSF-TransUNet, is proposed, which incorporates a Feature Fusion Attention Block (FFA-Block) to enhance the fusion of multi-scale features. This approach facilitates dense feature interactions through the integration of uniform attention, achieving this with minimal computational overhead. The experimental results on the Synapse and ACDC medical image segmentation datasets reveal that MSF-TransUNet outperforms existing models. Specifically, the average Hausdorff Distance (HD) on the Synapse dataset is reduced to 22.40 mm, accompanied by an impressive Dice Similarity Coefficient (DSC) of 80.78%. Furthermore, the model achieves a DSC of 91.52% on the ACDC dataset, demonstrating its superior performance. These findings highlight the potential of MSF-TransUNet in advancing medical image segmentation by effectively addressing the challenges of multi-scale feature fusion and semantic consistency.

1. INTRODUCTION

In addition to its many other uses, medical imaging segmentation can help with clinical diagnosis by making it easier to find diseased abnormalities and identify organs more accurately. Through this method, physicians can conduct qualitative and quantitative examinations on abnormal tissue and clinically relevant regions, enhancing the reliability and accuracy of clinical diagnoses. Tumor segmentation is especially crucial in surgical planning, which helps in the detection of precise tumor boundaries and guiding the surgical process. Hence, there is a clear need for advanced segmentation technologies in medical diagnostics.

Applications in segmentation tasks for medical imaging in recent years have been significantly influenced by Convolutional Neural Networks (CNNs). Models built on CNNs have been extensively developed due to their exceptional performance and straightforward network architecture. The FCN network [1], devoid of fully connected layers, employs convolutional layers in encoding and decoding stages. This design facilitates the preservation and reconstruction of spatial features, resulting in a streamlined model with enhanced generalization performance. The widely used U-Net [2] incorporates long skip connections at every level of its symmetric U-shaped encoder and decoder structure, aiming to mitigate the loss of spatial information caused by down-sampling operations. Despite the advantages of skip connections, U-Net still encounters challenges in modeling global multi-scale context and addressing the semantic gap.

UNet++ [3], based on nested U-Nets with dense skip connections and deep supervision methods, was developed to reduce the semantic gap between the encoder and decoder. U-Net 3+ [4] employs full-scale skip connections and deep supervision to learn multi-level features from feature maps that are aggregated at full scale. To address the narrow receptive field of CNNs, dilated convolutions are being used by DeepLab [5] to widen the receptive field. SegNet [6] enhances the up-sampling methods employed by the decoder used for features that are of low resolution while decreasing the quantity of trainable parameters and inference time. Utilizing a hybrid of the U-Net model and the atrous spatial pyramid pooling [7], DoubleU-Net [8] is able to precisely gather both spatial and contextual information. While U-Net structures have dominated image segmentation, they face challenges in acquiring long-term dependencies.

Transformers [9], initially proposed for natural language processing, excel at capturing long-range dependencies. Vision Transformers (ViTs) [10] subsequently applied

Transformers to computer vision, achieving performance comparable to convolution-based approaches. Multi-head self-attention is often credited with the superior performance of ViTs, as it facilitates global dependencies across every layer of the ViT architecture [11]. Swin Transformer [12] uses multiple layers of Transformers, employing shifted windows to ensure local information interacts effectively. Swin-UNet [13] made progress by combining Swin Transformer and U-Net architectures, but suffered from increased computational cost and limited performance improvement. Despite employing a pre-trained MaxViT encoder and a Convolutional Attention Mixing (CAM) decoder to enhance segmentation precision, MIST [14] faces challenges in capturing local pixel-level contexts. The Pyramid ViT [15] provides a convolution-free alternative with effective global context modeling. However, its high computational demands and dependence on large annotated datasets limit its applicability.

Numerous endeavors have been undertaken to integrate the transformer structure with U-Net. Taking TransUNet [16] for an example, it was one of the early efforts to leverage the advantages of ViTs, which also utilizing U-Net based frameworks s to enhance medical image segmentation performance. DA-TransUNet [17] incorporates dual attention block into TransUNet framework. AE-TransUNet+ [18] utilize CBAM [19] and depth-wise separable convolution (DSC) built upon the TransUNet architecture. CBAM-TransUNet [20] integrates the convolutional attention module within the bottleneck layer based on Swin Transformer. While attention techniques have been embedded into U-Net, Transformers, TransUNet, and other widely-used frameworks, for high-resolution images, the computational cost of self-attention can be very high, thereby further enhancement is necessary to optimize performance.

The skip connection serves as a crucial element in contemporary CNN architectures. Short skip connections in [21, 22] offer an additional pathway for uninterrupted gradient propagation. Long skip connections in studies [1, 2, 16, 23] preserve fine-grained features by linking earlier and deeper layers. Despite their application for amalgamating features through various pathways, the fusion of connected features typically involves addition or concatenation, allocating fixed weights to features regardless of content variations. Meanwhile, low-level features, like contours and edge, are essential for precise segmentation because of their ability to preserve key details. Nonetheless, their efficacy declines when they traverse several intermediary levels. However, low-level features, simply integrated with high-level feature as in Swin-UNet, TransUNet and DA-TransUNet, may lead to inconsistent spatial alignment, as high-level features capture a broader receptive field while low-level features preserve local fine details. This challenge underscores the critical need for effective multi-scale feature fusion in complex segmentation task for medical image.

In summary, balancing segmentation performance and computational efficiency remains a key challenge in designing medical image segmentation models based on deep learning. An effective multi-scale feature fusion strategy is essential for improving segmentation performance. Despite the fact that models such as TransUNet, which combine U-Net and Transformer, have been enhanced, they are still haunted by some intrinsic deficiencies:

(1) The extensive use of transformer blocks significantly increases the number of parameters and computational costs, but not necessarily with proportional improvements in segmentation accuracy.

(2) The researchers observed that ViT-based model attention maps favor dense interactions over sparse ones. This preference holds even though dense attention maps are more difficult to learn [11].

(3) Most existing methods struggle to fully utilize information at different scales. The challenge of balancing detailed features from low-level layer and semantic representations from high-level layer limits their ability to address scale discrepancies in medical image segmentation.

MSF-TransUNet, a novel TransUNet-based segmentation network, is introduced in this paper as a solution to aforementioned constraints. It enhances multi-scale feature fusion without significantly compromising computational efficiency. MSF-TransUNet introduces a novel FFA-Block for dynamically fusing detailed and semantic features with learned spatial attention mechanisms. With this, the model can address semantic gaps in the encode-decoder architecture while effectively utilizing information at various scales. By fusing local fine details and global context selectively, MSF-TransUNet enhances segmentation performance with improved feature representation.

Furthermore, to enhance global and local feature interactions, MSF-TransUNet integrates a large kernel convolutional modulation ($\geq 7\times7$) for suppressing redundant information and preserving valuable details for segmentation. Moreover, this paper integrates Context Broadcasting (CB) [11] into the Multilayer Perceptron Layer (MLP) so that all tokens can interact with global contextual information, thereby augmenting performance with merely linear computational cost increase, demonstrating a clear advantage over Transformer models, which typically suffer from quadratic complexity increase.

To evaluate MSF-TransUNet, this paper performs comprehensive experiments on two benchmark medical datasets for segmentation tasks: Synapse and ACDC. The findings indicate that MSF-TransUNet significantly surpasses current leading methods in both segmentation accuracy and computational efficiency. The primary contributions of this paper are highlighted below:

(1) MSF-TransUNet, a novel TransUNet-based segmentation framework, is introduced in this paper, which enhances multi-scale feature fusion and gains better segmentation performance in medical imaging applications.

(2) An innovative FFA-Block is presented to address semantic inconsistency problem more effectively in encoder-decoder structure for improving multi-scale feature fusion.

(3) Incorporating the CB module improves dense feature interactions in MSF-TransUNet, making both local and global feature learning more effective with minimal added computational complexity.

(4) MSF-TransUNet exhibits outstanding results on two medical imaging datasets, Synapse and ACDC, showcasing its superiority over current Transformer models.

2. METHODS AND MATERIALS

Figure 1 shows the MSF-TransUNet architecture with three components: encoder, skip connection, and decoder. (1) The encoder learns global features by dividing the provided medical images into 2D patches, flattening them into sequences and using the self-attention mechanism of Transformer model. Besides, CB module enhances feature

learning and model generalization. (2) Skip connections are redesigned in this new architecture by combining the low-dimensional features with the original skip connection layers, hence enhancing the decoder's capacity to include multi-scale features. Rather than directly adding features from a single encoder level to the decoder, the model makes use of multiple levels' information, which is obtained from the newly designed skip connections, through FFA-Block to enrich the learned representations. This ensures that the new skip connections' effectiveness by selectively enhancing and refining features obtained. Such an approach also enables the decoder to balance global contextual information with local fine details more efficiently. (3) The decoder consists of three core parts: Feature Concatenation, the FFA-block and up-sampling operations. The FFA-block, comprised of a Pre-Fusion block, Feature Fusion block, and Feature Selection block, which could adaptively grasp the most informative features, hence improving segmentation accuracy.

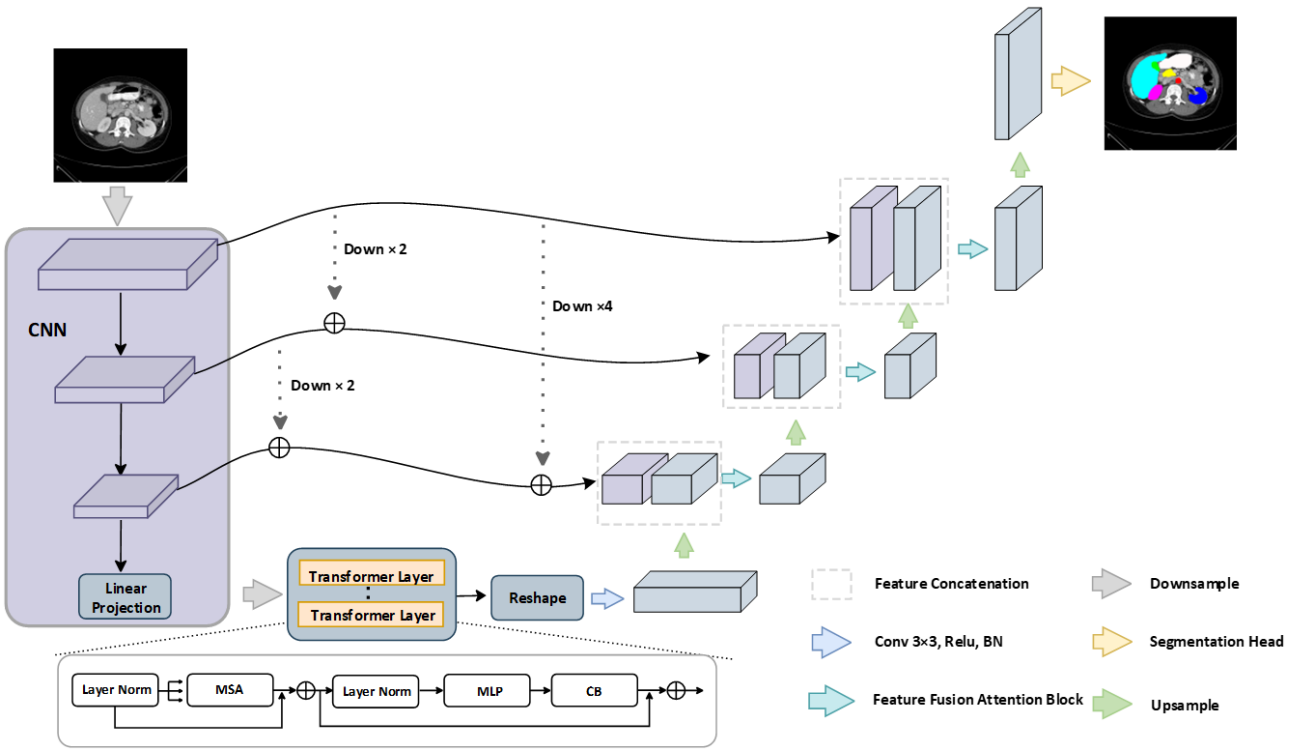

Figure 1. The structure of proposed MSF-TransUNet

2.1 Encoder incorporating CB module

MSF-TransUNet, like TransUNet [16], employs a hybrid architecture combining CNNs and Transformers in its encoder. It also incorporates the CB module [11], which enhances the model by promoting dense interactions. Previous research has shown the benefits of adding extra dense interactions in ViTs. Given the inherent challenge in learning dense attention through gradient descent, researchers manually implemented this process using a straightforward yet effective module called CB. The CB module is smoothly incorporated into the MLP layers of the MSF-TransUNet encoder, achieved with just a single line of code: X = 0.5 * X + 0.5 * X.mean(dim=1, keepdim=True) [11]. This CB module, by introducing uniform attention, inserts the token resulting from average pooling to each token individually, for infusing dense interactions. This integration not only simplifies the overall optimization process for MSF-TransUNet but also enhances its generalization. The CB module redirect MSF-TransUNet's focus from modeling dense attention maps to acquiring other valuable information, incurring only negligible additional operations for both inference and training.

2.2 Decoder

The architecture of the MSF-TransUNet decoder is akin to that of the TransUNet decoder. As shown in Figure 1, the MSF-TransUNet decoder utilizes three main operations: up-sampling operations, feature concatenation, and the FFA-Block. The model's final layer is a segmentation head that generates a feature map representing the prediction results.

2.2.1 FFA-block

To leverage features of varied scales extracted from the CNN part of the encoder and minimize the semantic disparity within the encoder-decoder architecture, optimizing the integration process of multi-scale features becomes imperative. In this paper, an FFA-Block, as illustrated in Figure 2, is introduced. The FFA-Block consists of a Pre-Fusion Block, a Feature Fusion Block, and a Feature Selection Block. This design aims to bridge the semantic gap effectively by utilizing multi-scale information while enhancing attention to capture salient features.

Specifically, this paper utilizes skip connections with elaborate FFA-Block to merges multiple low-dimensional features with the ones obtained at the previous decoding stage, which could attain a comprehensive hierarchical feature map. This map is subsequently incorporated into a Pre-Fusion Block for initial feature extraction, followed by a Feature Fusion Block that assigns a unique spatial importance map for channels, which could direct the model to prioritize crucial regions within every channel [24], ensuring the interaction of information across different scales.

The Feature Selection Block generates a more efficient

spatial attention map leveraging a convolutional modulation operation [25] after passing through two consecutive 3×3 convolutional layers. This block selects task-relevant features and filters out low-frequency information regions, which could augment the model's capacity to extract meaningful information.

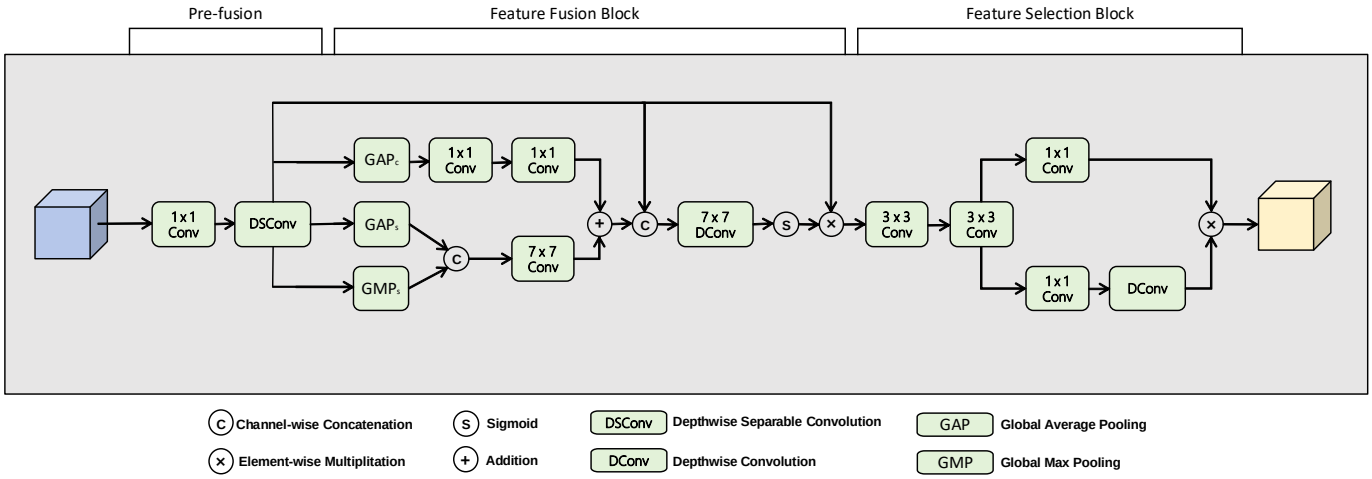

Figure 2. The schematic of FFA-block

2.2.2 Pre-fusion block

The Pre-Fusion block inputs from the previous decoder section and the CNN component within the encoder via skip connections. In this paper, Pointwise Convolution (*PWConv*) is chosen as the channel context aggregator, effectively combining information across channels and enabling the model to extract rich representational details from the input feature representations. After being processed by *PWConv*, these features are passed to DSC, which first applies a depth-wise convolution (*DConv*) independently to each individual channel, after which comes a $PWConv$ that transforms the output channels of the $DConv$ to a new channel space [26]. This structure allows for better fusion of features with inconsistent semantics and the extraction of spatial features from the input feature map while preserving channel information. Given $X \in R^{C \times H \times W}$, where X is the input with C channels and a size of $H \times W$, $F(X) \in R^{C \times H \times W}$ is computed as follows:

$$F(X) = B\left(PWConv(X)(\delta)\left(B\left(DConv\left(\delta\left(B(PWConv(x))\right)\right)\right)\right)\right) \quad (1)$$

Here, δ denotes ReLU6 activation function, B denotes batch normalization, $PWConv$ denotes pointwise convolution, $DConv$ denotes depthwise convolution and $F(X)$ denotes the feature map obtained from the concatenated features via Pre-Fusion block.

2.2.3 Feature fusion block

As shown in following equations, the feature map F, obtained from Pre-Fusion block, which contains multi-scale information, is subsequently processed by the Feature Fusion Block. To obtain a channel-specific spatial importance map [24], spatial global average pooling ($\mathrm{GAP_s}$) and spatial global max pooling ($\mathrm{GMP_s}$) are applied on spatial dimension of feature map $F \in \mathbb{R}^{C \times H \times W}$ to produce $F_{avg}^{s} \in \mathbb{R}^{1 \times H \times W}$ and $F_{max}^{s} \in \mathbb{R}^{1 \times H \times W}$. These operations extract spatial-wise feature information to encode critical spatial dependencies and discriminative patterns. Simultaneously, using global average pooling ($\mathrm{GAP_c}$) across the channel dimension, a global channel descriptor $F_{avg}^{c} \in \mathbb{R}^{C \times 1 \times 1}$ is generated from the feature map F.

$$F_{avg}^{s} = GAP_s(F) \quad (2)$$

$$F_{max}^{s} = GMP_s(F) \quad (3)$$

$$F_{avg}^{c} = GAP_c(F) = \frac{1}{H \times W}\sum_{h=1}^{H}\sum_{w=1}^{W} F(h,w) \quad (4)$$

GAP_c denotes global average pooling across the spatial dimension. GMP_s denotes global max pooling along the spatial dimension, highlighting the most salient spatial feature per channel. GAP_c denotes global average pooling across the channel dimension. For every position (h, w), the feature values from all channels are averaged.

By concatenating F_{avg}^{s} *and* F_{max}^{s} along the channel dimension to form a spatial descriptor and performing a convolution with 7×7 kernel size, more contextual information is captured, resulting in a spatial importance map M_s. Channel attention map is then obtained using Eq. (6), following the approach of SENet [27], where the core process involves 1×1 convolutions for channel compression and expansion, which adjusts the weight coefficients of each channel, thereby enhancing feature representation capabilities.

Additionally, residual connections address the vanishing gradient issue by incorporating the previous input feature F directly into later layers. By fusing M_c and M_s through element-wise summation, the semantic content of the input feature F serves as a guiding signal to derive the final channel-wise spatial importance map, ensuring alignment with the original feature structure. Finally, the fused features F' are computed using Eq. (7). This fusion approach achieves a dual contribution: it dynamically generates distinct spatial importance map per channel to prioritize critical regions, while simultaneously retaining fine-grained details from shallow layers and semantic context from deeper layers. The synergy

of these properties ensures precise segmentation of anatomical structures in medical imaging by balancing semantic and detailed features.

$$M_s = Conv_{7\times7}[F^s_{avg}, F^s_{max}] \tag{5}$$

$$M_c = Conv_{1\times1}\left(\delta\left(Conv_{1\times1}\left(F^c_{avg}\right)\right)\right) \tag{6}$$

$$F' = F \odot \sigma\left(DConv_{7\times7}([F, (M_s + M_c)])\right) \tag{7}$$

δ is the ReLU activation function, while $Conv_{k\times k}$ represents a convolution with a $k\times k$ kernel, $[\cdot]$ represents channel-wise concatenation. To optimize computational efficiency and minimize model complexity, the architecture employs a bottleneck design (Eq. (6)), a $Conv_{1\times1}$ first compresses the channel dimension and a subsequent $Conv_{1\times1}$ restores the channel dimension from C/r to C. In this paper, r is empirically set to 8. $DConv_{7\times7}$ denotes a 7×7 depth-wise convolution to capture spatial dependencies. $\odot$ denotes the Hadamard (element-wise) product, and σ represents the Sigmoid activation function.

2.2.4 Feature selection block

After features pass through Pre-Fusion block and Feature Fusion Block, they are forwarded to two standard convolution layers with a 3×3 kernel. Subsequently, a convolutional modulation (ConvMod) is employed to enhance spatial encoding efficiency within the Transformer U-Net framework. ConvMod is achieved by utilizing large kernels (≥7×7) within depth-wise convolutional layers [25, 28, 29], demonstrated in Eqs. (10)-(12). It facilitates task-relevant feature selection and attenuates redundant low-frequency information from the feature map F'.

Self-attention [9] takes a $X \in \mathbb{R}^{N\times C}$, where $N=H\times W$ denotes the flattened spatial dimensions (height H, width W), and C represents the channel depth. The feature map X is first mapped through three learnable linear transformations to derive the key (K), query (Q), and value (V) matrices, each with dimensions $\mathbb{R}^{N\times C}$. The self-attention output is calculated as a weighted sum of the value matrix V, with the weights are determined by a normalized similarity score matrix A. This score matrix A is derived from the pairwise interactions between Q and K, capturing long-range dependencies across the input sequence.

$$Attention(Q, K, V) = A \cdot V \tag{8}$$

$$A = \frac{\mathrm{Softmax}(\mathrm{QK^T})}{\sqrt{d}} \tag{9}$$

where, d is the dimension of Q and K vectors, serving as a scaling factor. The Softmax function normalizes the similarity scores into a probability distribution across the input sequence. The attention mechanism computes the output as context-aware aggregation of the V matrix, where the aggregation weights are determined by A. These weights encode pairwise affinities between Q and K, enabling the model to prioritize semantically relevant regions of the input.

Within ConvMod module, instead of computing the self-attention similarity matrix A as in traditional ViTs, self-attention is approximated by replacing matrix multiplications with the Hadamard product combined with depth-wise convolution operations to compute the output. To be specific, given $X \in \mathbb{R}^{H\times W\times C}$, the ConvMod module applies a simple depth-wise convolution with a $K\times K$ kernel to capture spatial patterns. An element-wise Hadamard product operation is subsequently performed to compute the output Z. The process is formally expressed as follows:

$$Z = A \odot \mathrm{V} \tag{10}$$

$$A = DConv_{k\times k}(W_1 X) \tag{11}$$

$$V = W_2 X \tag{12}$$

In those equations, $\odot$ represents the Hadamard product. W_1 and W_2 denote the weight matrices of two linear layers, while $DConv_{k\times k}$ refers to a depthwise convolution with a $K\times K$ kernel. This formulation allows each spatial location (h, w) to capture correlations with all pixels in its surrounding $K\times K$ receptive field, ensuring effective local and global feature interactions.

Specifically, the Feature Selection block in MSF-TransUNet employs ConvMod with a 7×7 kernel size following the deepest skip connection, while the Feature Selection Blocks at other stages utilize a 11×11 kernel size. ConvMod streamlines feature selection by replacing the computationally expensive self-attention mechanism with a combination of depth-wise convolutions and the Hadamard product, significantly reducing computational overhead while preserving spatial dependencies.

As shown in Table 1, the computational complexity of ConvMod is $O(N\cdot C)$, whereas the complexity of self-attention is $O(N^2\cdot C)$. This quadratic complexity in self-attention becomes prohibitive when processing high-resolution medical images. In contrast, ConvMod provides a more scalable and efficient alternative, ensuring computational feasibility without sacrificing performance. By leveraging large kernels in a computationally efficient manner, ConvMod improves the extraction of features at various scales, rendering it especially effective for medical image segmentation tasks.

Table 1. Complexity of self-attention and ConvMod

Layer Type	Complexity
Self-Attention	$O(N^2\cdot C)$
ConvMod	$O(N\cdot C)$

2.3 Dataset and evaluation

This paper utilized the Synapse [30] and ACDC [31] datasets. The Synapse dataset comprises 30 CT-enhanced abdominal scans, each containing a varying number of slices. This dataset includes eight abdominal organs with corresponding labels. The Synapse dataset could be accessed through https://www.synapse.org/Synapse:syn3193805/wiki/217789.

The ACDC dataset consists of cardiac MRI scans from 100 patients, with labels for the left ventricle (LV), right ventricle (RV), and myocardium (MYO). The data splitting for both datasets follow the approach outlined in study [32]. The ACDC dataset could be accessed through https://www.creatis.insa-lyon.fr/Challenge/acdc/.

For the evaluation of our approach's segmentation results, this paper employed DSC [33] and HD [34]. The corresponding equations are presented in Eq. (13) and Eq. (14):

$$DSC(M,N) = 2\frac{|M \cap N|}{|M| + |N|} \tag{13}$$

$$HD(M,N) = \max\big(h(M,N), h(N,M)\big) \tag{14}$$

Here, *M* and *N* represent the point sets corresponding to the Ground Truth and the predicted values, respectively. The term *h*(*M*,*N*) refers to the supremum of the minimum distances from all points in set *M* to set *N*, as shown in Eq. (15). This metric quantifies the largest deviation between the two sets.

$$h(M,N) = \max_{m \in M} \min_{n \in N} \|m - n\| \tag{15}$$

2.4 Loss function

To optimize the effectiveness of the segmentation approach, a composite loss function which incorporates both Cross-Entropy Loss and Dice Loss is implemented. The formulation is as follows:

$$Loss = 0.4 \times Cross\text{-}Entropy\ Loss + 0.6 \times Dice\ Loss \tag{16}$$

$$Cross\text{-}Entropy\ Loss = -\sum_{i=1}^{n} y_i\ log\ p_i \tag{17}$$

$$Dice\ Loss = 1 - \frac{2|A \cap B|}{|A| + |B|} \tag{18}$$

Here, *n* represents the total amount of pixels from an input image, y_i denotes the ground-truth for the *i*-th pixel and p_i is the predicted probability for the same pixel. The sets *A* and *B* are the point sets for the Ground Truth and the predicted segmentation, correspondingly.

The designed composite loss function integrates the merits of both loss functions. Cross-Entropy Loss has a good performance in classifying each pixel independently, and therefore the predicted probabilities closely match the actual class labels. Dice Loss, on the other hand, focuses on where the ground truth and the expected segmentations overlap, which is particularly beneficial for addressing class imbalance and improving boundary precision of segmentation. By employing these two loss functions, the model strikes a compromise between accurate pixel classification and well-defined boundaries and therefore attains better overall segmentation performance.

2.5 Implementation details

The MSF-TransUNet model was implemented using Python 3.9, PyTorch 2.0.0, and CUDA 11.8 with an NVIDIA GeForce 4090 GPU. The images were pre-resized to 224×224 pixel. The model parameters were initialized by the pre-trained "R50-ViT" model. For the models trained with the SGD optimizer, a weight decay of 0.0001 was applied, and training was conducted for 150 epochs. The momentum was set to 0.9, and an initial learning rate of 0.01. The learning rate had a dynamic decay schedule, which was set as:

$$lr = base_{lr} \times (1.0 - \frac{iter\_num}{max\_iterations})^{0.9} \tag{19}$$

where, *iter_num* denotes the current iteration, *max_iterations* represents the total number of iterations and $base_{lr}$ is the initial learning rate. This decay schedule enables a gradual reduction of the learning rate, which helps stabilize the optimization and promotes convergence.

MSF-TransUNet uses techniques for data enhancement commonly employed in TransUNet, such as random rotation, horizontal and vertical flipping, and spatial rescaling to the desired resolution, to increase the model's generalizability. The augmentations allow the model to support diversity in medical images and enhance the model's cross-dataset robustness.

3. RESULTS

3.1 Experiment results on Synapse dataset

This paper presents a comparison of MSF-TransUNet on Synapse dataset. Table 2 presents a comprehensive comparison of the segmentation accuracy of various methods, measured by average DSC in percentage and average HD in millimeters. The proposed MSF-TransUNet is highlighted with the highest average DSC of 80.78% and the lowest HD of 22.40 mm among all models provided in Table 2. It indicates that MSF-TransUNet produces the best overall segmentation outcome with the most precise boundaries. Compared with the baseline model (TransUNet), MSF-TransUNet shows a 3.46% improvement in DSC and a decrease of 8.23 mm in HD.

In the segmentation results, MSF-TransUNet outperforms TransUNet across various organs: 0.67% for the aorta, 10.11% for the gallbladder, 0.68% for the left kidney, 2.46% for the right kidney, 0.43% for the liver, 4.74% for the pancreas, 4.4% for the spleen, and 4.19% for the stomach. Overall, both the average DSC and the DSC for all eight organs were significantly higher than the baseline model, indicating enhanced segmentation performance by MSF-TransUNet

Table 2. Segmentation accuracy of different methods on Synapse dataset

Model	DSC	HD	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
U-Net [2]	75.43	37.72	87.07	60.01	78.53	73.11	92.73	56.23	85.77	69.99
U-Net++(resnet-50) [3]	76.13	32.16	85.95	58.63	79.26	73.81	93.77	57.35	85.10	75.16
TransUNet [16]	77.32	30.63	87.44	61.22	81.33	76.66	94.40	57.23	84.68	75.61
UCTransNet [35]	77.54	33.93	87.86	63.99	83.12	77.15	93.54	54.49	85.24	74.96
Swin-UNet [13]	76.04	26.49	83.27	64.06	79.54	73.42	93.12	55.76	86.05	73.13
VA-TransUNet [32]	79.21	28.32	87.72	66.96	82.22	74.54	94.45	60.34	88.37	79.03
DA-TransUNet[17]	79.85	23.72	86.57	61.30	83.73	80.84	94.59	59.86	89.60	82.35
MSF-TransUNet	**80.78**	**22.40**	**88.11**	**71.33**	82.01	79.12	**94.83**	**61.97**	89.08	79.80

Figure 3 illustrates the visualization of the segmentation results. The hybrid Transformer and CNN architecture, combined with the FFA-Block and CB module, facilitates learning dense spatial interactions and capturing useful information more adequately, which enables MSF-TransUNet to extract image features more effectively, resulting in enhanced segmentation performance.

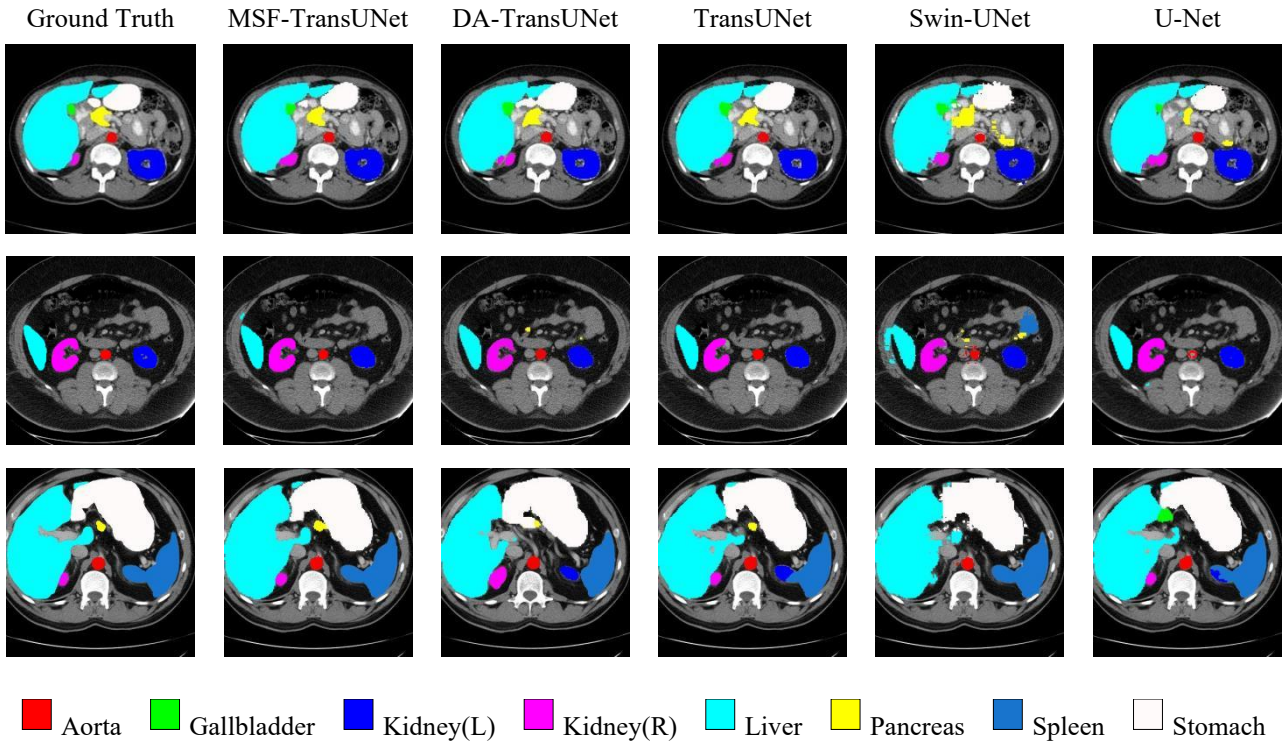

Figure 3. Qualitative comparison of various methods through visualization. From left to right: (a) Ground Truth, (b) MSF-TransUNet, (c)DA-TransUNet (d)TransUNet, (e) Swin-UNet, (f) U-Net. MSF-TransUNet produces fewer false positives and preserves finer details

3.2 Experiment results on ACDC dataset

Likewise, MSF-TransUNet was employed on the ACDC dataset, with performance measured by the average DSC in percentage for RV, MYO, and LV. The results are presented in Table 3. MSF-TransUNet demonstrates the best overall performance with an average DSC of 91.52%. It achieves the highest DSC for RV at 90.25% and LV at 96.59%, along with a strong performance for MYO at 87.72%, which surpasses most state-of-the-art segmentation approaches.

Table 3. Segmentation accuracy of different methods on ACDC dataset

Model	Year	DSC	RV	MYO	LV
U-net [2]	2015	89.18	85.71	86.01	95.83
U-Net++(resnet-50) [3]	2018	89.61	88.36	84.84	95.64
TransUNet [16]	2021	90.08	88.27	85.86	96.10
UCTransNet [35]	2022	89.72	87.52	85.71	95.94
Swin-UNet [13]	2022	88.65	86.73	83.77	95.46
VA-TransUNet [32]	2022	91.00	88.88	87.78	96.35
HiFormer-B [36]	2023	89.06	87.28	84.52	95.37
DA-TransUNet [17]	2024	91.09	89.16	87.69	96.43
MSF-TransUNet	2025	**91.52**	**90.25**	87.72	**96.59**

These impressive results highlight MSF-TransUNet's ability to accurately segment different cardiac structures, showcasing its robustness and effectiveness. For the RV, MSF-TransUNet's 90.25% DSC indicates that it performs particularly well at identifying the structural details and boundaries, much better than previous models. Similarly, its highest DSC for LV at 96.59% indicates superb accuracy in delineating the LV, which is critical to making accurate cardiac studies and interventions. The strong performance in MYO segmentation, with a DSC of 87.72%, also signifies the capacity of the model to deal with the complexity of myocardial tissue, which is generally difficult due to its variable and complex nature.

3.3 Ablation study

3.3.1 Effect of CB module

In order to evaluate the performance of the CB module, we made experiments by removing the CB module from the MLP layers of Transformers of MSF-TransUNet and adding the CB module to the baseline model of TransUNet for comparison. Experimental results on Synapse dataset are presented in Table 4. The DSC% values are superior for methods with CB module, indicating that the CB module broadcasts contextual information well.

3.3.2 Effect of FFA-block

To validate the efficacy of FFA-Block, we conducted an ablation study by removing it from MSF-TransUNet and adding FFA Block into the baseline model of TransUNet for comparison, the results are showed in Table 4. We observed a distinct enhancement in both experiment with FFA block, which indicated that FFA block excel at fusing multi-scale features and capturing useful information from them.

Table 4. Ablation study on the impact of CB and FFA module (%)

Method	DSC	HD	FLOPs (G)	Params (M)	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
Baseline (TransUNet)	77.32	30.63	24.728	93.232	87.44	61.22	81.33	76.66	94.40	57.23	84.68	75.61
Baseline+FFA	78.37	29.92	33.614	98.701	88.04	66.19	80.10	75.26	94.31	61.02	87.04	75.03
Baseline+CB	78.18	28.67	24.728	93.232	87.45	63.93	81.52	75.39	94.05	57.83	87.78	77.49
Baseline+FFA+CB	80.78	22.40	33.614	98.701	88.11	71.33	82.01	79.12	94.83	61.97	89.08	79.80

4. DISCUSSIONS

4.1 Discussions on synapse dataset

This paper provides a comprehensive comparison performance of MSF-TransUNet, TransUNet, DA-TransUNet, Swin-UNet, and U-Net on the Synapse dataset with qualitative comparison illustrated in Figure 3 and quantitative comparison showed in Table 2. The results, illustrated in Figure 3, lead to several important conclusions. (1) Compared to other Transformer-based and hybrid models, models based entirely on CNNs tend to produce more mis-segmentation issues. Specifically, as shown in the third row of Figure 3, U-Net misclassifies the spleen, while in the first row, it produces false positives for the pancreas. This indicates that other Transformer-based and hybrid models have stronger global context encoding and semantic differentiation capabilities. (2) The results in the third row of Figure 3 indicate that MSF-TransUNet preserves the overall shape of the organs more precisely compared to other approaches and effectively reducing mis-segmentation. In contrast, U-Net, TransUNet, and DA-TransUNet exhibit varying levels of misclassification in the spleen segmentation, while Swin-UNet shows coarse segmentation boundaries and false-positive regions. These results reflect that MSF-TransUNet achieves superior accuracy and robustness in segmentation.

Experiments show that MSF-TransUNet performs better in segmentation tasks, preserving shape information and effectively representing global context as well as fine-grained information. This is evidenced by the observation that MSF-TransUNet performs better in most organs with the highest DSC values for the aorta (88.11%), gallbladder (71.33%), liver (94.83%), and pancreas (61.97%), and its performance on the kidneys and spleen remain stable. To ensure the statistical significance of the enhanced performance, the baseline model and MSF-TransUNet were compared through a paired t-test on the Synapse datasets. The results confirmed a statistically significant improvement in Dice score across the Synapse dataset ($t = 3.031$, $p = 0.0191$). This overall improvement reflects the success of MSF-TransUNet in effectively integrating global and local feature, making it especially robust for challenging segmentation tasks of medical imaging.

4.2 Discussions on ACDC dataset

MSF-TransUNet was applied to the segmentation task on ACDC dataset with performance measured using average DSC in percentage for RV, MYO, and LV. The experimental results, as shown in Table 3, indicate that MSF-TransUNet performs the best with an average DSC of 91.52%, surpassing other models like DA-TransUNet with 91.09% and TransUNet with 90.08%. It achieves the highest DSC for RV at 90.25% and LV at 96.59%, and demonstrates strong performance for MYO with 87.72%, which is better than the performance of most existing segmentation methods.

These results prove the effectiveness and strength of MSF-TransUNet in accurately segmenting different cardiac structures. In the case of the RV, the DSC of 90.25%, outperforming other models like VA-TransUNet (88.88%) and TransUNet (88.27%), indicates that MSF-TransUNet excels in identifying the structural edges and details of the RV, much better than the previous models. Similarly, the highest DSC for the LV also demonstrates excellent accuracy in delineating the LV, which is crucial for accurate cardiac assessments and treatments. The high performance in MYO segmentation also reflects the model's ability to deal with the complexities of myocardial tissue, which is typically difficult to segment due to its complicated and highly variable nature.

4.3 Further work

The outstanding performance of MSF-TransUNet on the Synapse and ACDC datasets can be attributed to its hybrid Transformer and CNN structure. The structure effectively integrates global context with local feature. The incorporation of FFA-block and CB module allows the model to learn dense spatial interactions and extract key information more effectively, leading to more accurate and robust segmentation results.

Despite these advantages are observed, MSF-TransUNet faces some challenges. Firstly, the addition of FFA-blocks comes with a computational cost. Nevertheless, this increase in complexity is mitigated by the use of convolutional modulation (ConvMod) instead of self-attention, which effectively limits the overall computational burden. Further optimizations are still needed for real-time applications, particularly in low-resource environments. Future research will explore techniques such as model pruning and lightweight attention mechanisms to further reduce computational costs while preserving segmentation accuracy. Additionally, this present work addresses two-dimensional medical image segmentation, but most medical imaging modalities like CT and MRI provide three-dimensional volumetric data. In order to fill this gap, future work will focus on extending MSF-TransUNet to volumetric 3D medical image segmentation, assessing its effectiveness in volumetric scenarios. Moreover, future efforts will aim to optimize the model for real-time applications as well as further enhance its capabilities to three-dimensional medical image segmentation to leverage its full potential in clinical practice.

5. CONCLUSIONS

This paper introduces MSF-TransUNet, a TransUNet-based neural network, to enhance the accuracy of medical image segmentation. Extensive experiments on multi-organ segmentation (Synapse) and cardiac segmentation (ACDC) tasks indicate that MSF-TransUNet outperforms state-of-the-art approaches in both performance and generalizability, efficiently enhancing the segmentation accuracy.

With relatively low extra computational cost, the inclusion

of the CB module provides the dense interactions required by the model, further boosting its performance. The proposed FFA-Block effectively mitigates semantic gaps in the encoder-decoder framework. By generating channel-specific spatial weight in the Feature Fusion Block and extracting useful information through ConvMod in the Feature Selection Block, the FFA-Block is capable of focusing on large-scale global targets as well as localized fine-grained detail. This improves the model's accuracy as well as its capacity for generalization across various medical imaging segmentation tasks. In summary, MSF-TransUNet achieves state-of-the-art performance in medical image segmentation by combining the low-complexity CB module and the FFA module, which effectively leverages multi-scale features information. Future work will extend this architecture to 3D medical image analysis, exploring volumetric extensions of CB and FFA modules to tackle challenges in multi-modal 3D segmentation, and evaluating scalability on large-scale clinical datasets.

ACKNOWLEDGEMENTS

This research is supported by the Open Research Fund of State Key Laboratory of Digital Medical Engineering (Grant No.: 2024-M10); Natural Science Foundation of Nanjing University of Posts and Teclecommunications (Grant No.: NY223093); Natural Science Foundation of Zhejiang Province (Grant No.: LGG22F020040); Wenzhou scientific research project (Grant No.: ZG2024013).

REFERENCES

[1] Long, J., Shelhamer, E., Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, pp. 3431-3440. https://doi.org/10.1109/CVPR.2015.7298965

[2] Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In Medical image computing and computer-assisted intervention–MICCAI 2015: 18th International Conference, Munich, Germany, pp. 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[3] Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J. (2018). Unet++: A nested U-Net architecture for medical image segmentation. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, pp. 3-11. https://doi.org/10.1007/978-3-030-00889-5_1

[4] Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Wu, J. (2020). Unet 3+: A full-scale connected unet for medical image segmentation. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, pp. 1055-1059. https://doi.org/10.1109/ICASSP40776.2020.9053405

[5] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4): 834-848. https://doi.org/10.1109/tpami.2017.2699184

[6] Badrinarayanan, V., Kendall, A., Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(12): 2481-2495. https://doi.org/10.1109/TPAMI.2016.2644615

[7] Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, pp. 801-818. https://doi.org/10.1007/978-3-030-01234-2_49

[8] Jha, D., Riegler, M.A., Johansen, D., Halvorsen, P., Johansen, H.D. (2020). DoubleU-Net: A deep convolutional neural network for medical image segmentation. In 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), Rochester, MN, USA, pp. 558-564. https://doi.org/10.1109/CBMS49503.2020.00111

[9] Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I. (2017). Attention is all you need. Neural Information Processing Systems, 2017: 30.

[10] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[11] Hyeon-Woo, N., Yu-Ji, K., Heo, B., Han, D., Oh, S.J., Oh, T.H. (2023). Scratching visual transformer's back with uniform attention. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, pp. 5807-5818. https://doi.org/10.1109/ICCV51070.2023.00534

[12] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, pp. 10012-10022. https://doi.org/10.1109/ICCV48922.2021.00986

[13] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M. (2022). Swin-unet: Unet-like pure transformer for medical image segmentation. In European Conference on Computer Vision, Israel, pp. 205-218. https://doi.org/10.1007/978-3-031-25066-8_9

[14] Rahman, M.M., Shokouhmand, S., Bhatt, S., Faezipour, M. (2024). Mist: Medical image segmentation transformer with convolutional attention mixing (cam) decoder. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, pp. 404-413. https://doi.org/10.1109/wacv57701.2024.00047

[15] Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Shao, L. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, pp. 568-578. https://doi.org/10.1109/iccv48922.2021.00061

[16] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y.,

Zhou, Y. (2021). Transunet: Transformers make strong encoders for medical image segmentation. arXiv Preprint arXiv:2102.04306. https://doi.org/10.48550/arXiv.2102.04306

[17] Sun, G., Pan, Y., Kong, W., Xu, Z., Ma, J., Racharak, T., Xin, J. (2024). DA-TransUNet: Integrating spatial and channel dual attention with transformer U-Net for medical image segmentation. Frontiers in Bioengineering and Biotechnology, 12: 1398237. https://doi.org/10.3389/fbioe.2024.1398237

[18] Jia, Y., Su, Z., Wan, G., Liu, L., Liu, J. (2023). AE-TransUNet+: An enhanced hybrid transformer network for detection of lunar south small craters in LRO NAC images. IEEE Geoscience and Remote Sensing Letters, 20: 1-5. https://doi.org/10.1109/LGRS.2023.3294500

[19] Woo, S., Park, J., Lee, J.Y., Kweon, I.S. (2018). Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, pp. 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[20] Chen, X., Yang, L. (2022). Brain tumor segmentation based on CBAM-TransUNet. In Proceedings of the 1st ACM Workshop on Mobile and Wireless Sensing for Smart Healthcare, New York, NY, United States, pp. 33-38. https://doi.org/10.1145/3556551.3561192

[21] Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., De Lange, T., Halvorsen, P., Johansen, H.D. (2019). Resunet++: An advanced architecture for medical image segmentation. In 2019 IEEE International Symposium on Multimedia (ISM), San Diego, CA, USA, pp. 225-2255. https://doi.org/10.1109/ism46123.2019.00049

[22] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, pp. 770-778. https://doi.org/10.1109/CVPR.2016.90

[23] Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S. (2017). Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, pp. 2117-2125. https://doi.org/10.1109/CVPR.2017.106

[24] Chen, Z., He, Z., Lu, Z.M. (2024). DEA-Net: Single image dehazing based on detail-enhanced convolution and content-guided attention. IEEE Transactions on Image Processing, 33: 1002-1015. https://doi.org/10.1109/tip.2024.3354108

[25] Hou, Q., Lu, C.Z., Cheng, M.M., Feng, J. (2024). Conv2former: A simple transformer-style convnet for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12): 8274-8283. https://doi.org/10.1109/tpami.2024.3401450

[26] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, pp. 1251-1258. https://doi.org/10.1109/CVPR.2017.195

[27] Hu, J., Shen, L., Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 7132-7141. https://doi.org/10.1109/CVPR.2018.00745

[28] Wang, Z., He, X., Li, Y., Chuai, Q. (2022). EmbedFormer: Embedded depth-wise convolution layer for token mixing. Sensors, 22(24): 9854. https://doi.org/10.3390/s22249854

[29] Ding, X., Zhang, X., Han, J., Ding, G. (2022). Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, pp. 11963-11975. https://doi.org/10.1109/CVPR52688.2022.01166

[30] Landman, B., Xu, Z., Igelsias, J.E., Styner, M., Langerak, T., Klein, A. (2015). Segmentation outside the cranial vault challenge. In MICCAI: Multi Atlas labeling Beyond Cranial Vault-Workshop Challenge.

[31] Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Jodoin, P.M. (2018). Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?. IEEE Transactions on Medical Imaging, 37(11): 2514-2525. https://doi.org/10.1109/TMI.2018.2837502

[32] Jiang, T., Xu, T., Li, X. (2022). VA-TransUNet: A u-shaped medical image segmentation network with visual attention. In Proceedings of the 2022 11th International Conference on Computing and Pattern Recognition, pp. 128-135. https://doi.org/10.1145/3581807.3581826

[33] Asgari Taghanaki, S., Abhishek, K., Cohen, J.P., Cohen-Adad, J., Hamarneh, G. (2021). Deep semantic segmentation of natural and medical images: a review. Artificial Intelligence Review, 54: 137-178. https://doi.org/10.1007/s10462-020-09854-1

[34] Lalonde, M., Beaulieu, M., Gagnon, L. (2001). Fast and robust optic disc detection using pyramidal decomposition and Hausdorff-based template matching. IEEE Transactions on Medical Imaging, 20(11): 1193-1200. https://doi.org/10.1109/42.963823

[35] Wang, H., Cao, P., Wang, J., Zaiane, O.R. (2022). Uctransnet: rethinking the skip connections in U-Net from a channel-wise perspective with transformer. In Proceedings of the AAAI Conference on Artificial Intelligence, California, USA, pp. 2441-2449. https://doi.org/10.1609/aaai.v36i3.20144

[36] Heidari, M., Kazerouni, A., Soltany, M., Azad, R., Aghdam, E.K., Cohen-Adad, J., Merhof, D. (2023). HiFormer: Hierarchical multi-scale representations using transformers for medical image segmentation. In 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, pp. 6202-6212. https://doi.org/10.1109/WACV56688.2023.00614