POLITECNICO DI TORINO
Repository ISTITUZIONALE

MultiTrans: Multi-branch transformer network for medical image segmentation

Original
MultiTrans: Multi-branch transformer network for medical image segmentation / Zhang, Y., Balestra, G., Zhang, K.e., Wang, J., Rosati, S., Giannini, V.. - In: COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE. - ISSN 0169-2607. - ELETTRONICO. - 254:(2024). [10.1016/j.cmpb.2024.108280]

Availability:
This version is available at: 11583/2989563 since: 2024-06-19T15:55:11Z

Publisher:
Elsevier

Published
DOI:10.1016/j.cmpb.2024.108280

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
Elsevier postprint/Author's Accepted Manuscript

© 2024. This manuscript version is made available under the CC-BY-NC-ND 4.0 license
http://creativecommons.org/licenses/by-nc-nd/4.0/.The final authenticated version is available online at:
http://dx.doi.org/10.1016/j.cmpb.2024.108280

(Article begins on next page)

01 November 2026

MultiTrans: Multi-Branch Transformer Network for Medical Image Segmentation

Yanhua Zhang[a,b], Gabriella Balestra[a], Ke Zhang[b,*], Jingyu Wang[b], Samanta Rosati[a], Valentina Giannini[c,d]

[a]*Department of Electronics and Telecommunications, Politecnico di Torino, Corso Duca degli Abruzzi 24, Turin, 10129, Italy*
[b]*School of Astronautics, Northwestern Polytechnical University, 127 West Youyi Road, Xi'an, 710072, China*
[c]*Surgical Sciences Department, University of Turin, Turin, 10124, Italy*
[d]*Radiology Unit, Candiolo Cancer Institute, FPO-IRCCS, Candiolo, 10060, Italy*

Abstract

Background and Objective: Transformer, which is notable for its ability of global context modeling, has been used to remedy the shortcomings of Convolutional neural networks (CNN) and break its dominance in medical image segmentation. However, the self-attention module is both memory and computational inefficient, so many methods have to build their Transformer branch upon largely downsampled feature maps or adopt the tokenized image patches to fit their model into accessible GPUs. This patch-wise operation restricts the network in extracting pixel-level intrinsic structural or dependencies inside each patch, hurting the performance of pixel-level classification tasks.

Methods: To tackle these issues, we propose a memory- and computation-efficient self-attention module to enable reasoning on relatively high-resolution features, promoting the efficiency of learning global information while effective grasping fine spatial details. Furthermore, we design a novel Multi-Branch Transformer architecture to provide hierarchical features for handling objects with variable shapes and sizes in medical images. By building four parallel Transformer branches on different levels of CNN, our hybrid network aggregates both multi-scale global contexts and multi-scale local features.

Results: MultiTrans achieves the highest segmentation accuracy on three medical image datasets with different modalities: Synapse, ACDC and M&Ms. Compared to the standard Self-Attention, the proposed efficient Self-Attention can largely reduce the training memory and computational complexity while even slightly improve the accuracy. Specifically, the training memory cost, FLOPs and Params of our efficient Self-Attention are 18.77%, 20.68% and 74.07% of the standard Self-Attention.

Conclusions: Experiments on three medical image datasets demonstrate the generality and robustness of our proposed network. The ablation study shows the efficiency and effectiveness of our proposed efficient Self-Attention. Code is available at: https://github.com/Yanhua-Zhang/MultiTrans-extension.

Keywords:
Medical Image Segmentation, Abdominal Multi-Organ Segmentation, Cardiac Segmentation, Deep Learning, Efficient Self-Attention, Parallel Transformer Branches.

1. Introduction

Medical image segmentation has been successive used in a wide range of medical applications based on various imaging modalities [1, 2, 3, 4, 5]. However, visual inspection of these images is often time-consuming, highly subjective and relays on experienced operators [6, 7]. Convolutional neural networks (CNN) are widely used in computer-aided medical image segmentation and have shown remarkable progress [8, 9, 10, 11]. U-Net [10] proposes a symmetric encoder-decoder architecture to recover detailed information and aggregate multi-level features, becoming a paradigm for medical image segmentation. A series of followers (e.g., U-Net++ [12], UNet3+ [13], ResUNet [14], etc) also adopt this encoder-decoder frameworks, showing excellent representation ability. To further enhance the performance of CNN, NAG-Net [15] proposes two UNet-based nested sub-networks to focus on the clinician's visually salient region, and BDNet [16] adds a boundary refinement module to learn accurate boundary locations.

However, CNN-based networks exhibit limitations in modeling long-range dependency due to the intrinsic locality properties of convolution kernel [17, 18]. Transformer, which has recently been exploited in computer vision, is notable for its ability of modeling long-range dependence. Also, some efforts have been made in medical image segmentation to use it to compensate for the shortcomings of CNN [19, 20, 21, 22]. Among them, TransUNet [21] puts ViT on the top of CNN to extract global context information, which is the first work to leverage the power of Transformer to segment medical images. TransFuse [20] combines the multi-level global information extracted by one single Transformer branch with the multi-level local information provided by CNN branch at each individual level. These methods prove the effectiveness of Transformer in providing global context information to segment organs or

*Corresponding author
Email addresses: yanhua.zhang@studenti.polito.it (Yanhua Zhang), gabriella.balestra@polito.it (Gabriella Balestra), zhangke@nwpu.edu.cn (Ke Zhang), jywang@nwpu.edu.cn (Jingyu Wang), samanta.rosati@polito.it (Samanta Rosati), valentina.giannini@unito.it (Valentina Giannini)

tumors from a medical scan.

Despite the advances, Transformer are also known for their memory and computational inefficient computation, which is the square of the number of pixels. This makes it impossible to directly apply standard self-attention in large spatial resolution feature maps. To fit their model into easily accessible GPUs, many methods choose to inference on largely downsampled feature maps (e.g., the last stage feature of CNN) [23, 24, 25] or adopt the tokenized image patches [26, 18, 20, 27]. This may not maximize the advantages of the Transformer. The detailed information is lost due to the downsampling operation (pooling and stride layers) in CNN, and the patch-wise operation restricts the network in obtaining pixel-level intrinsic structural or dependencies inside each patch.

To reduce the model complexity of Transformer, several Efficient Self-Attention (ESA) methods have been adopted in medical image segmentation. Inspired by the low-rank theory in Linformer [28], TransHRNet [29] uses the strided convolution to reduce the sequence length of Key and Value to save memory and computational costs. CoTr [30] adopts Deformable Self-Attention [31] to select representative feature points from input sequences to reduce computational complexity, while Valanarasu et al [32] employs the axial attention [33] as its ESA by calculating affinity matrix independently along the height and width axes of feature maps. However, the point-selection [34] or various downsampling operations [35, 36, 37] in the low-rank ESA pose a risk of losing important information, and the receptive field of axial attention [38] is limited to a certain axis. In addition, all the above methods do not consider the cost of linear projection and value transformation operations of SSA, which can be a big burden when the input features have large channel dimensions. Transformer-based methods await a better solution to promote the efficiency of learning global information while holding an effective grasp of spatial details. To tackle these problems, without following the above low-rank designs, we propose a hybrid ESA to reduce memory and computational complexities. Firstly, when computing the affinity matrix, we change the multiplication order in adherence to the associativity law of matrix product, reducing the quadratic complexity relative to input size to linear, which is mathematically equivalent to standard self-attention [39]. Secondly, we further reduce the memory usage by introducing the Head-Sharing operation, that is, multiply heads use the same affinity matrix for transformation. Finally, by using the same linear layer for Key and Value, we employ this sharing mechanism to reduce the cost of projection operations. Our main motivation for proposing ESA is to enable the Transformer branch to reason on large-resolution resolution feature maps to reduce the loss of fine spatial details for precise segmentation.

In addition, medical images are characterized by high intra-class and inter-class variation. For example, organs, tumors and lesion regions appear in different shapes and sizes across patients. Therefore, capturing multi-scale information is very important for segmenting objects with multiply scales. However, most existing methods only use one Transformer branch to extract the single-scale global feature while ignore the multi-scale information, resulting in low segmentation accuracy [21, 40, 23]. For handling this problem, DS-TransUNet [27] uses two swin Transformers with different patch size to obtain both coarse and fine-grained feature representations. UTNet [37] applies low-rank efficient self-attention module to each level of the encoder to obtain multi-scale global context information. The above methods still leave space for improving the effectiveness of capturing and fusing multi-scale features. DS-TransUNet uses two pure Transformers to extract multi-scale global representations without integrating local features from CNN, and these two kinds of information are complementary [20, 40, 22]. In addition, we point out that the UTNet architectur designed for multi-scale global feature extraction is a sub-optimal choice for two main reasons: firstly, the self-attention layers used directly on the low-level layers of CNN are too shallow to effectively learn long-range dependence; Secondly, the convolutional layers in the decoder part of UTNet may blur the global features obtained from the Transformer, as the square convolution kernels can merge information from the surrounding feature pixels.

To this end, taking full advantage of CNN's inherent feature hierarchy, we build four parallel Transformer branches on different levels of the CNN to explicitly obtain multi-scale global contexts and fuse them with local features extracted by convolutional layers though long skip connections. Moreover, we design a top-down path to flow the high-level features extracted by deep convolutional layers of CNN to lower levels to provide semantic cues for Transformer branches, especially the low-level branches. When fusing local and global features, as pointed by [20, 41], the low-level CNN features could be noisy, so we modify the attention gate module [41] to use global features from Transformer to guide the filtering of local features and enhance local details. Finally, instead of using the U-shaped decoder, we directly use large-scale upsampling to fuse features from Transformer branches to avoid global features being blurred by convolution kernels.

In a nutshell, we have four main contributions.

- We propose a hybrid ESA to reduce the memory and computational complexity of the three main steps of SSA. When computing the affinity matrix, we change the matrix multiplication order to reduce the complexity from quadratic to linear of the sequence length, which is mathematically equivalent to SSA, thus avoiding information loss caused by the downsampling operations in low-rank ESA. In addition, we propose two sharing mechanisms to further reduce the cost of linear projection and value transformation operations in SSA.

- To effectively learn the long-range dependence at each scale, we build independent parallel Transformer branches on different levels of the CNN to explicitly extract multi-scale global features, and design a top-down path to provide high-level semantic features extracted by deep layers of CNN to the learning of low-level Transformer branches.

- Without following the classical U-Net like architecture, we use large-scale upsampling to fuse features from multi-branch Transformers to avoid global contexts being

blurred by convolutional kernels. Furthermore, the attention gated module is adopted to use global contextual information from Transformer to filter local features of CNN and guide the local-global feature fusion on each individual branch.

- To boost network training, we add the In-deep Supervision (IDS) customized for the parallel Transformer branches architecture to improve gradient flow and enhance feature representation.

2. Related Work

In this section, we mainly review Transformer-based networks and efficient self-attention methods in the field of medical image segmentation.

2.1. Pure/Hybrid Transformer Networks

Transformer-based segmentation networks can be roughly divided into two categories: pure Transformers [42, 43, 44, 27] and hybrid Transformers [21, 20, 22, 35, 37]. Swin-Unet [42], which combines the Swin Transformer block with the classic U-shaped architecture design, is the first 2D pure Transformer model proposed in medical domain, achieving improved results compared to CNN networks. D-Former [44] performs self-attention within each patch to calculate local context representation, and proposes dilated self-attention inspired by dilated convolution to learn long-range dependencies. Then, they arrange these two attention mechanisms alternately to build a pure Transformer for 3D medical image segmentation.

However, due to the lack of spatial induction bias, pure Transformers exhibit limited ability in extracting localized information [21] and cannot perform well on small-scale datasets [37]. Similar to TransFuse, CTC-Net [22] also builds a dual encoding path, using parallel CNN and Transformer branches to extract local context and global features respectively. MISSFormer [35] injects locality into the Transformer by adding depth-wise convolution layers in the feed-forward network (FFN) of self-attention module, and uses repeated layer norm to re-integrate the local and global features. These methods only use one Transformer branch to extract the single-scale global feature while ignore the multi-scale representations, leading to low segmentation accuracy [21, 40, 23]. UTNet [37] follows the standard design of UNet, but replace the last convolution layer of each stage of the CNN encoder with Transformer modules to obtain multi-scale global context information. Here, we point out that the UTNet architecture design leave space for improvement for two main reasons: firstly, the shallow self-attention layers cannot effectively learn long-range dependence at each scale, especially at the low-level layers of CNN; Secondly, the convolution-based decoder part of UTNet is a sub-optimal choice for fusing global features, as the square convolution kernels can merge information from the surrounding feature pixels to re-extract local information, thereby blurring the global features obtained from the Transformer.

To effectively learn the long-range dependence at each scale, we build independent parallel Transformer branches on different levels of the CNN to explicitly extract multiscale global features, and design a top-down path to provide high-level semantics to the learning of low-level Transformer branches. Instead of using the classical U-Net like architecture to fuse features from multi-branch Transformers, we use large-scale upsampling to avoid global contexts being blurred by convolutional kernels.

2.2. Efficient Self-Attention

Standard Self-Attention (SSA) suffers from quadratic computational and spatial complexity in terms of image pixels, making it unsuitable for high-resolution feature maps. Medical image segmentation is a position-sensitive task, so high-resolution features play an important role for segmenting boundary and small objects, which are easily lost due to largely downsampling or patch-wise operations. To solve the above problem, a large number of Efficient Self-Attention (ESA) modules are adopted in medical image segmentation. From the theory in Linformer [28], self-attention is low-rank for long sequences, indicating that most information is highly redundant. Inspired by this finding, most methods adopt various down-sampling operations to reduce the feature spatial size of Key and Value. These downsampling operations can be average/max pooling [35], linear projection [36], strided convolution [29], or bilinear interpolation [37]. CoTr [30] and MCTrans [34] use Deformable mechanism proposed by [31] to select a fixed number of elements from the sequence to reduce computational complexity. Without following the low-rank ESA design, Valanarasu et al. [32] and Wang et al. [38] applied the axial attention [33] to perform self-attention independently on the height and width axis of feature maps. The downsampling operation in low-rank ESA and points selection operation of Deformable attention could lead to information loss, while the receptive field of axial attention is limited to a certain axis. In addition, all the above methods only focus on saving the cost of the affinity matrix calculation, while ignoring the burden of linear projection and value transformation operations of SSA.

Different from these methods, as shown in Fig. 1(c), we propose a hybrid ESA module that combines the Order-Changing operation with two types of sharing mechanisms to reduce the costs of the three main steps of SSA (affinity matrix calculation, linear projections and value transformation), respectively. In particular, the Order-Changing operation follows the associativity law of matrix product, which is mathematically equivalent to SSA, avoiding the risk of information loss caused by the downsampling or point selection operations in other ESA methods.

3. Method

3.1. Overview

Fig. 1 shows the architecture of our multi-branch Transformer network, and illustrates the difference between the SSA and the proposed ESA. The ensuing subsections will first

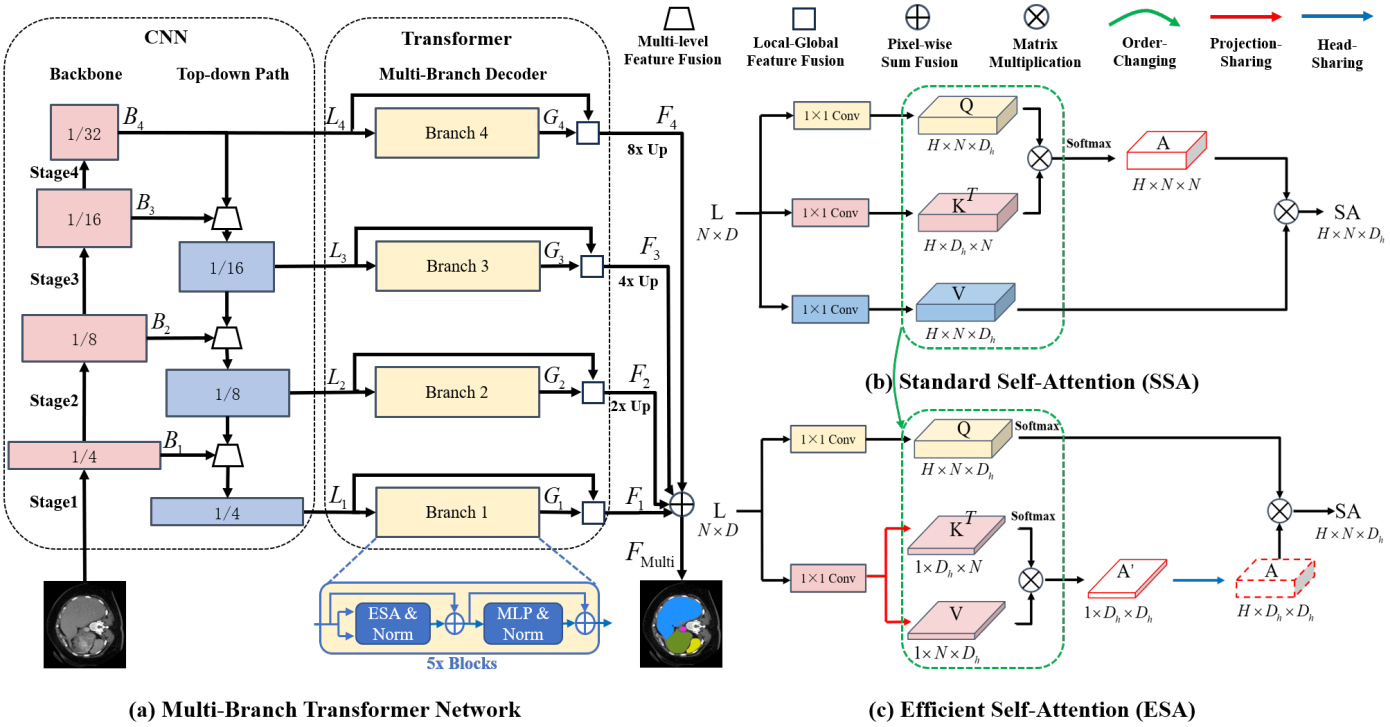


Figure 1: Overview of our proposed hybrid network, and the comparison between the standard Self-Attention and the proposed efficient Self-Attention.

overview the theory of SSA, which consists of three main steps: three linear projections, computing affinity matrix, and Value transformation. Then, to reduce the memory or computational cost of the above steps, we combine the Order-Changing with Head- and Projection-Sharing operations to build our hybrid ESA, and provide detailed complexity analysis of each operation. In the following Section. 3.4, we sequentially introduce the detailed design of CNN part, parallel Transformer architecture and the local-global feature fusion module. Furthermore, we analyze the difference of our MultiTrans with a contemporary work, and compare different architecture designs shown in Fig. 3 to give a clear understanding of the motivations of our network. Finally, we present a special In-deep Supervision (IDS) to supervise the training of our hybrid network.

3.2. Standard Self-Attention

A standard multi-head self-attention (MSA) [39] can be expressed as a flattened feature map $L \in \mathbb{R}^{N\times D}$ (N means the number of pixels and D expresses the feature dimension) passes through H separate parallel self-attention heads. The h-th head can be formally written as:

1. Uses three different linear projections to compute the Query, Key, and Value:

$$\begin{aligned} Q^h &= w_Q^h \mathrm{Norm}(L), \\ K^h &= w_K^h \mathrm{Norm}(L), \\ V^h &= w_V^h \mathrm{Norm}(L). \end{aligned} \tag{1}$$

Here, $Q, K, V \in \mathbb{R}^{N\times D_h}$, and we set the feature dimension of them equal to D_h for simplicity. $w_Q, w_K, w_v \in \mathbb{R}^{D_h\times D}$ are trainable weights of linear layers. Norm means layer normalization.

2. Computes the affinity matrix A^h scaled by $\frac{1}{\sqrt{D_h}}$ and transformed value SA^h:

$$A^h = \mathrm{Softmax}\left(\frac{Q^h (K^h)^T}{\sqrt{D_h}}\right), \quad \mathrm{SA}^h = A^h \cdot V^h. \tag{2}$$

3. Concatenates the outputs of the H heads and reprojects back onto $\mathbb{R}^D$.

$$\mathrm{MSA} = W_{\mathrm{Project}}\left[\mathrm{SA}^1; \ldots; \mathrm{SA}^H\right]^T, \tag{3}$$

where $W_{\mathrm{Project}} \in \mathbb{R}^{D\times D_h H}$. Subsequent residual connections, ReLU activations, layer normalization and the feed-forward network are omitted for brevity.

3.3. Efficient Self-Attention

The dot-product in MSA (Eq. (2)) leads to $O(HD_hN^2)$ computational complexity and the resulting affinity matrix has $O(HN^2)$ memory complexity. Normally, N is much larger than D_h especially for feature maps from very shallow layers, thus the number of pixels dominates the self-attention computation and memory costs. Since Eq. (2) can be regarded as a series of matrix multiplication, we change the order of computation to reduce memory and computation [45, 46]. We also remove the scaling factor $\frac{1}{\sqrt{D_h}}$ because we find from the results in Table 7 that it has a negative impact on our ESA. As shown in Fig. 1(c), we can rewrite Eq. (2) as:

$$\begin{aligned} A^h &= \mathrm{Softmax}\left(K^h\right)^T V^h, \\ \mathrm{SA}^h &= \mathrm{Softmax}\left(Q^h\right) A^h. \end{aligned} \tag{4}$$

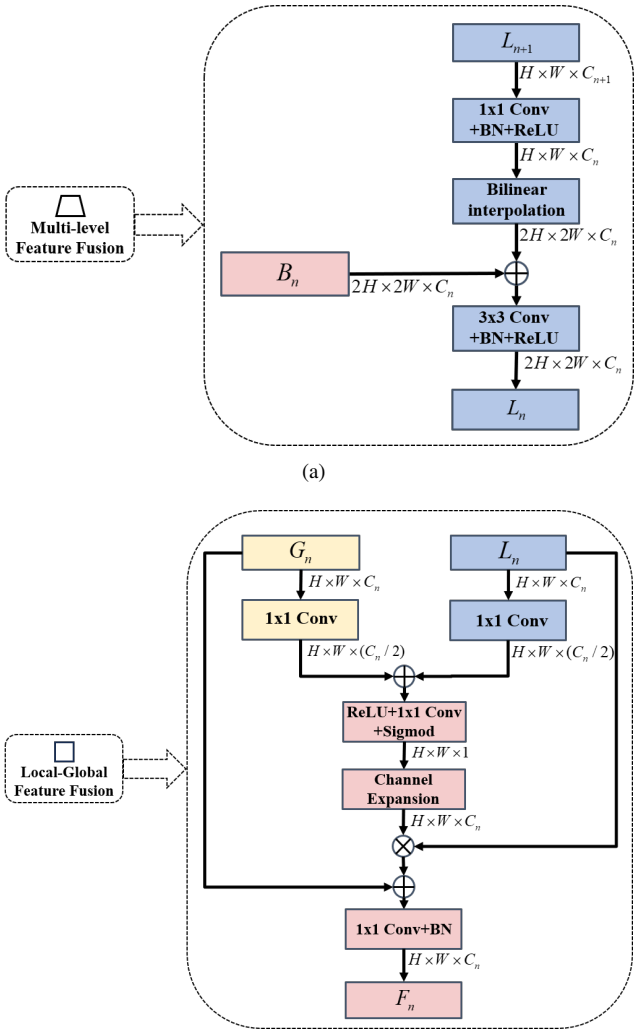


Figure 2: Detailed design of the multi-level feature fusion module in the Top-down path of the CNN part and the Local-global feature fusion module.

This simple operation can largely reduce the memory and computational complexity of the affinity matrix to $O(HD_h{}^2)$ and $O(HD_h{}^2N)$. In order to further reduce the memory usage of the affinity matrix, we introduced the Head-Sharing mechanism, that is, multiply heads use the same affinity matrix for transformation, and can be formulated as:

$$\begin{aligned} A_{\text{Share}} &= \text{Softmax}\left(K^T\right)V, \\ \text{SA}^h &= \text{Softmax}\left(Q^h\right)A_{\text{Share}}. \end{aligned} \tag{5}$$

By combining with Order-Changing operation, this results in $O(D_h{}^2)$ memory complexity and $O(D_h{}^2N)$ computational complexity, further being reduced by 1/H.

In addition, the total memory and computation of the three linear projections in Eq. (1) are $3HD_hDN$ and $3HD_hN$, respectively, which are non-negligible compared to the dot-product in our efficient self-attention. Therefore, we also employ the sharing mechanism for the projections, which has been proved its effectiveness in NLP tasks [28]. As shown in Fig. 1(c), this is easily achieved by using the same linear layer to go from L to K and V, and a separate one for Q. By combining with Head-Sharing mechanism, the memory and computation are decreased to $(H+1)D_hDN$ and $(H+1)D_hN$, which are $(H+1)/(3H)$ of the original projections.

3.4. Network Architecture

As shown in Fig. 1(a), we use ResNet-50 as the backbone for local feature extraction, which has the typical CNN structure of four residual convolution stages. By stacked stride layers and pooling layers, different stages of the backbone can provide multi-level local features ($B_1 \sim B_4$) with different receptive fields. Besides, the local inductive bias of convolutional layers can speed up the training process and avoid using large-scale pretraining [21, 25, 24].

In addition, upon multi-level features of the backbone, we build a top-down path to flow the high-level features to the low-level features. The CNN part of our network is similar to the UNet architecture. Differently, we adopt the top-down path to provide high-level semantic cues to the low-level Transformer branches, while UNet aims to use high-resolution features to progressively recover the spatial details of the topmost feature. Instead of using deconvolution to restore spatial resolution, we use simple bilinear upsampling followed by 3x3 convolution to fuse multi-level features. The details of multi-level feature fusion module are shown in Fig. 2(a), which can be expressed mathematically as:

$$L_n = \text{Conv}_{3\times3}(\text{Up}_{2\times}(\text{Conv}_{1\times1}(L_{n+1})) + B_n), n = 3, 2, 1, \tag{6}$$

where $L_4 = B_4$ and $\text{Up}_{2\times}$ represents $2\times$ upsampling via bilinear interpolation.

As for decoder, we use our proposed Efficient Self-attention module to build four parallel Transformer branches on different levels of CNN to acquire multi-scale global features ($G_1 \sim G_4$). Low-level features of CNN have larger spatial resolution with retaining finer local information. Therefore, inferring on these features to obtain high-resolution context information helps segment small-scale objects and detailed architectures.

On each branch, we use a long skip connection to fuse the local feature with the global representation, since these two kinds of information are complementary [40]. If directly using the sum operation for feature fusion, the i-th branch feature is as follows:

$$F_i = \text{Conv}_{1\times1}(L_i + G_i). \tag{7}$$

However, as low-level CNN features could be noisy [20], this simple sum operation brings noise or irrelevant information into global features, reducing the effectiveness of feature fusion. Thus, as shown in Fig. 2(b), we modify the attention gate module proposed by [41] to enhance local details and suppress irrelevant regions in local features. By adopting the spatial attention mechanism, we use global features G_i to guide the filtering of local features L_i, and Eq. (7) can be rewritten as:

$$\begin{aligned} A_i &= \sigma_2(\text{Conv}_{1\times1}(\sigma_1(\text{Conv}_{1\times1}(L_i) + \text{Conv}_{1\times1}(G_i)))), \\ F_i &= \text{Conv}_{1\times1}(A_i * L_i + G_i), \end{aligned} \tag{8}$$

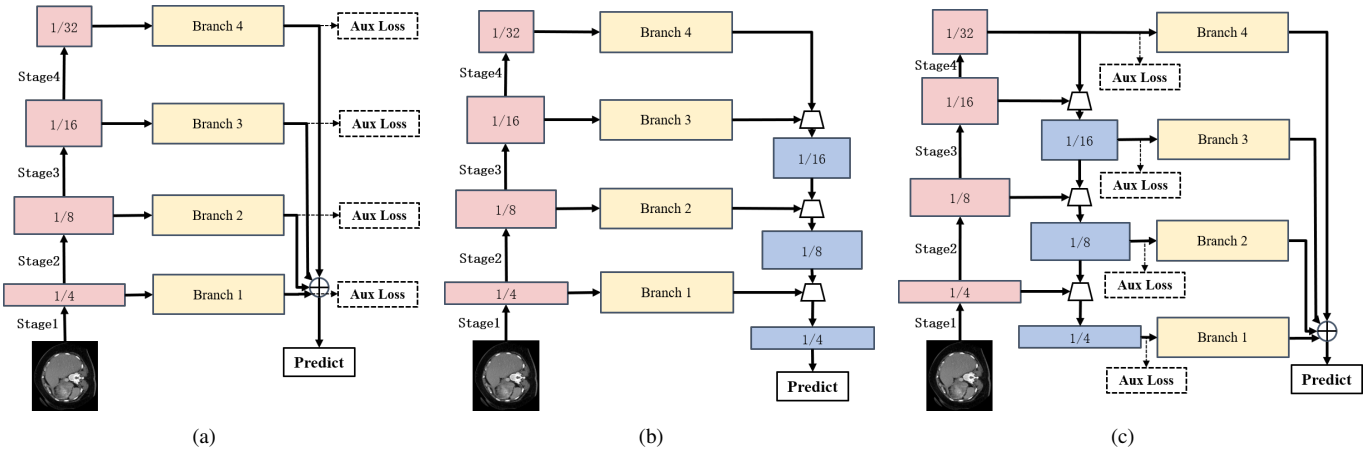


Figure 3: Comparison of different macro architecture designs. "Aux Loss" denotes auxiliary loss. The dashed arrows indicate the location of the auxiliary losses. (a) Our initial design and the deep supervision. (b) UNet-like architecture. (c) Our final design and the In-deep Supervision.

where σ_1 and σ_2 correspond to the ReLU activation function and the sigmoid normalization respectively. A_i is attention maps and * represents element-wise multiplication.

Finally, by using the bilinear interpolation upsampling and pixel-wise sum fusion, the final output F_{Multi} aggregated four branches' feature maps:

$$F_{\text{Multi}} = \text{Conv}_{1\times1}(F_1 + \text{Up}_{2\times}(F_2) + \text{Up}_{4\times}(F_3) + \text{Up}_{8\times}(F_4)). \tag{9}$$

3.5. Architecture Comparison and Analysis

The contemporary work TransHRNet [29] shares a similar motivation with us, but we have several notable differences with it: firstly, our Order-Changing operation is mathematically equivalent to standard self-attention [45] and the sharing mechanisms help us further improve efficiency and reduce redundant information, while TransHRNet uses strided convolution to reduce the spatial size of Key and Value, inevitably leading to information loss; Secondly, TransHRNet directly applies parallel interactive Transformer branches to each level of the CNN backbone, while we build four independent parallel Transformer branches upon the top-down path which provides hierarchical represents for low-level Transformer branches; Thirdly, TransHRNet follows the U-shaped architecture to use deconvolution in the decoder to restore resolution, while we directly use large-scale bilinear interpolation to upsample and fuse features from Transformer branches.

To give a clear understanding of the motivations of our design, we compare different macro architecture designs in Fig. 3. Our initial design aims to extract multi-scale global features through direct inference on different levels of backbone. We also illustrate a design of building the top-down path on the output of multiple Transformer branches, which follows a UNet-like architecture. The top-down path used in U-net architecture mainly enables precise localization by combing high resolution features from backbone [10]. Interestingly, we found that this design had a negative impact on segmentation performance compared to the initial design (Table 8). We attribute this to the convolutional layers in the top-down path blurring the global features obtained from the Transformer, as the square convolution kernels gathers information from the surrounding region, inevitably introducing irrelevant information from neighboring pixels. Differently, our final design uses the top-down path to fuse multi-level features of backbone to provide hierarchical feature cues for low-level Transformer branches, taking full advantage of CNN's inherent feature hierarchy. Besides, instead of using the U-shaped decoder, we directly use large-scale upsampling for fusing features from Transformer branches. Quantitative comparisons and corresponding analyses are given in Section. 5.4.3.

3.6. Objective Functions

Following [21, 37], we combine the Dice loss and Cross-Entropy loss to train the MultiTrans in an end-to-end manner, as below:

$$\mathcal{L}(G, S) = \phi_{CE}(G, S) + \phi_{DICE}(G, S). \tag{10}$$

Here, ϕ_{CE} is the Cross-Entropy loss function and ϕ_{DICE} denotes the Dice loss function. G means ground-truth. S expresses the predicted segmentation map generated from the segmentation head with one 1×1 convolutional layer.

Furthermore, as shown in Fig. 3(c), we add additional segmentation heads in front of the Transformer branches to compute auxiliary losses to improve gradient flow and enhance feature representation. We name it In-deep Supervision (IDS) to distinguish it from the deep supervision illustrated in Fig. 3(a). The In-deep Supervision can be written as:

$$\mathcal{L} = \mathcal{L}(G, S_{\text{Final}}) + \sum_{i=1}^{4} (\lambda_i \mathcal{L}(G, S_i)), \tag{11}$$

where S_{Final} and S_i are segmentation maps calculated from F_{Multi} and L_i. From λ_1 to λ_4 are hyperparameters that control the weight of auxiliary losses, and we experimentally set them to 0.4, 0.3, 0.2 and 0.1, respectively.

Table 1: Comparison with the state-of-art methods on the Synapse dataset. The highest and second highest accuracy of each organ are bolded and underlined, respectively. '$*$' and '†': Corresponding methods are reproduced by [21] and us, respectively.

Name	DSC	HD	DSC for each organ								Params
	(%) ↑	(mm) ↓	Ao	Ga	Ki(L)	Ki(R)	Li	Pa	Sp	St	(M) ↓
CNN-based:											
V-Net[9]	68.81	-	75.34	51.87	77.10	80.75	87.84	40.05	80.56	56.98	-
DARR[47]	69.77	-	74.74	53.77	72.31	73.24	94.08	54.18	89.90	45.96	-
R50+UNet* [10]	74.68	36.87	84.18	62.84	79.19	71.29	93.35	48.23	84.41	73.92	-
R50+Att-UNet* [41]	75.57	36.97	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95	-
Transformer/Hybrid:											
TransUNet[21]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62	105.3
UTNet† [37]	78.11	27.98	87.00	73.36	82.25	76.82	94.39	56.07	86.99	68.04	**9.5**
CTC–Net [22]	78.41	22.52	86.46	63.53	83.71	80.79	93.78	59.73	86.87	72.39	-
LeViT-UNet-384 [48]	78.53	16.84	87.33	62.23	84.61	80.25	93.11	59.07	88.86	72.76	52.2
SwinUnet [42]	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60	62.8
TransFuse-L† [20]	80.48	19.88	88.16	66.85	82.97	80.00	94.54	62.14	90.23	78.94	102.4
HiFormer-L [49]	80.69	19.14	87.03	68.61	84.23	78.37	94.07	60.77	90.44	82.03	29.5
MISSFormer [35]	81.96	18.20	86.99	68.65	85.21	82.00	94.41	65.67	91.92	80.81	42.5
MultiTrans	**84.82**	**12.66**	**88.66**	**76.50**	**86.48**	**84.34**	**94.79**	**69.81**	**93.11**	**84.89**	39.4

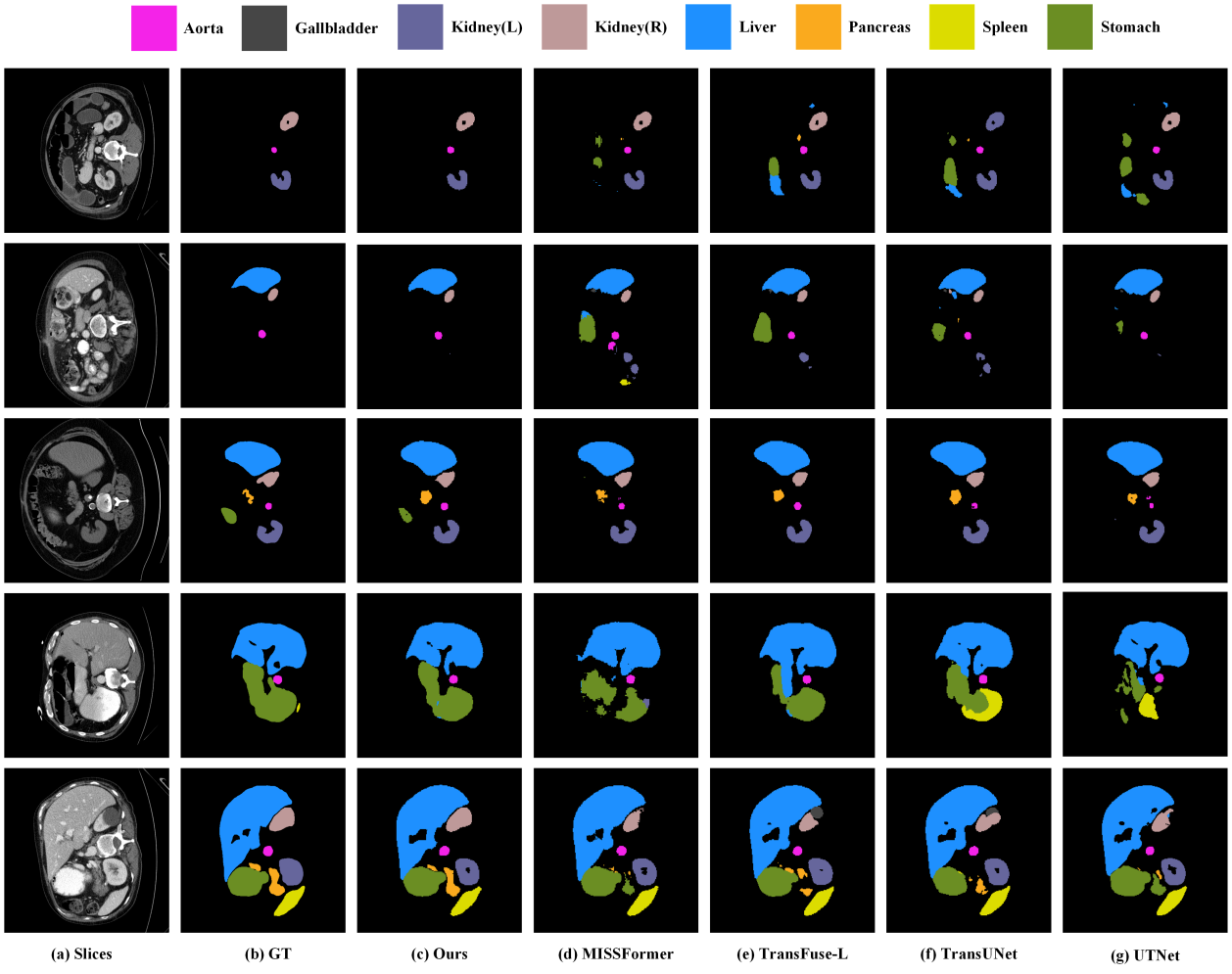


Figure 4: Qualitative results on Synapse dataset. Compared with other recently proposed Transformer-based methods, our network exhibits more accurate contours, higher intra-class consistency, and stronger inter-class discrimination ability.

4. Experiments

4.1. Dataset and Evaluation Metrics

4.1.1. Synapse

It is a widely used multi-organ segmentation dataset[1], which has 30 cases with 3779 axial abdominal clinical CT slices in total. Following [21, 47], 18 training cases (2212 axial slices) are random picked for training and 12 cases for validation. Each

[1]https://www.synapse.org/#!Synapse:syn3193805/wiki/217789

Table 2: Comparison with the state-of-art methods on the ACDC dataset. '∗' : Corresponding methods are reproduced by [50], and their backbones are based on ResNet-34 [51]. '†': Corresponding methods are based on 3D networks.

Name	DSC	Cardiac Structures			Params
	(%) ↑	RV	MYO	LV	(M)
CNN-Based:					
DeepLabv3+* [52]	88.25	85.41	85.44	93.90	26.3
PSPNet* [53]	88.75	85.99	86.39	93.87	**14.4**
Att-UNet* [41]	89.01	87.3	85.07	94.66	23.7
UNet* [10]	89.41	87.77	85.88	94.67	23.6
UNet++* [12]	89.58	87.23	87.13	94.37	24.4
PraNet [54]	90.16	87.21	88.73	94.54	-
nnUNet† [55]	91.61	90.24	89.24	95.36	30.8
Transformer/Hybrid:					
UNETR† [56]	88.61	85.29	86.52	94.02	92.5
TransUNet [21]	89.71	88.86	84.54	95.73	105.3
Swin-UNet [42]	90.00	88.55	85.62	95.83	62.8
LeViT-UNet-384 [48]	90.32	89.55	87.64	93.76	52.2
TransHRNet† [29]	91.00	90.29	86.88	95.82	36.8
MISSFormer [35]	91.19	89.85	88.38	95.34	42.5
nnFormer† [57]	92.06	90.94	89.58	95.65	149.6
D-Former† [44]	92.29	91.33	89.6	95.93	44.3
H2Former [50]	92.40	91.31	90.12	95.76	33.7
UNETR++† [36]	92.83	**91.89**	90.61	96.00	42.6
MultiTrans	**92.88**	91.33	**90.84**	**96.48**	39.4

Table 3: Comparison with the state-of-art methods on the M&Ms dataset. We not only reported the performance of all models trained and tested only on vendor A, but also provided the Dice accuracy of several networks trained on vendor A and B and tested on extra unseen vendors. '∗' and '†': Corresponding methods are reproduced by [37] and us, respectively.

Name	Only Vendor A		Seen		Unseen		Params
	DSC ↑	HD ↓	A	B	C	D	(M)
CNN-Based:							
UNet* [10]	86.4	13.9	-	-	-	-	**7.1**
ResUNet* [58]	86.9	11.48	87.9	87.6	85.7	84.2	9.4
Dual-Attn* [59]	87.0	11.3	88.0	88.1	85.8	84.4	9.7
CBAM* [60]	87.3	10.8	88.5	88.4	85.5	85.3	9.4
Transformer/Hybrid:							
TransUNet† [21]	87.65	11.22	87.7	87.7	84.9	86.0	105.3
SwinUNet† [42]	87.92	10.06	88.0	88.2	86.5	86.3	62.8
UTNet [37]	88.3	10.8	88.7	**88.7**	**86.6**	86.2	9.5
TransFuse-L† [20]	88.45	9.90	88.7	88.4	85.7	86.4	102.4
MultTrans	**88.81**	**9.09**	**89.1**	**88.7**	**86.6**	**86.7**	39.4

CT slice has 512 × 512 pixels, which are resized to 224 × 224 for training our network. We use the Dice-Similarity coefficient (DSC) to measure the overlapping between predictions and ground truth, and the 95% Hausdorff Distance (HD) to evaluate the quality of segmentation boundaries by calculating the maximum distance between the predicted boundaries and its ground truth. We also reported the DSC accuracy for each organ (Ao: Aorta, Ga: Gallbladder, Ki(L): Left Kidney, Ki(R): Right Kidney, Li: Liver, Pa: Pancreas, Sp: Spleen, St: Stomach).

4.1.2. ACDC

The Automated Cardiac Diagnosis Challenge[2] dataset consists of 100 cardiac MRI scans with manual annotations of the left ventricle (LV), the right ventricle (RV) and the myocardium (MYO). According to [21], 70, 10 and 20 scans are selected for training, validation and testing, respectively. All MRI slices are randomly croped to 224 × 224 for training. We provided the DSC for each cardiac structure and their average accuracy for comparison.

4.1.3. M&Ms

Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge [61] is another cardiac structure segmentation dataset that, like ACDC, also includes annotations for LV, RV, and MYO. Differently, we performed experiments on this dataset to measure the model robustness of MultiTrans.

[2]https://www.creatis.insa-lyon.fr/Challenge/acdc/

Following [37], first, 75 and 40 MRI scans from the same vendor A are used for training and testing. Second, 150 scans from vendors A (Siemens) and B (Philips) are used for training, and 200 scans from 4 different MRI vendors are used for testing, including the other 2 unseen vendors (C: GE, D: Canon). We randomly cropped each MRI slice to 256 × 256 for training and reported the DSC and 95% HD for evaluation.

4.2. Implementation Details

We conduct experiments based on PyTorch 2.0.1. The FLOPs and training memory are measured on one NVIDIA A800 (80GB). During training, we adopt the "poly" learning rate strategy by $1 - \left(\frac{iter}{\max_iter}\right)^{0.9}$ and the network is optimized using the Stochastic Gradient Descent (SGD) algorithm.

On the Synapse dataset, the initial learning rate, batch size and the epochs are set to 0.1, 24 and 150, respectively. Following the prior protocol [21, 42, 48, 49], we only use random rotation and flipping for data augmentation on Synapse. And, we do not adopt time-consuming evaluation tricks (e.g., flipping, gaussian noise and window sliding) to improve accuracy. Instead, we directly use resized images as input to do inference. We employ above simple training or evaluation strategies in Table 1 and Ablation Studies to show the performance improvements come from our architecture design, not from sophisticated training or evaluation tricks.

For the ACDC dataset, we set the initial learning rate, batch size and the epochs to 0.01, 24 and 200, respectively. Besides, to maximum the performance of our method and make a fair comparison with several 3D networks, following [55, 57, 56, 44, 36], we additional add scaling, Gamma transformation, Gaussian noise and mirroring in the data augmentation and adopt flipping, Gaussian noise and window sliding as the evaluation strategy.

When evaluating our method on M&Ms, the batch size and initial learning rate are 32 and 0.05 (0.1 when training with vendor A only). Consistent with the previous method [37], we also employ more data augmentation methods including scaling, brightness, Gaussian noise and Gamma transformation.

Table 4: Ablation experiments on the design of efficient self-attention. TM: Training Memory. '†': Calculated by the branch 1 with 5 Transformer layers and 8 heads.

OC	HS	PS	DSC	HD	TM (GB)	FLOPs† (G)	Params† (M)
✓	-	-	84.0 ± 0.5	13.4 ± 2.0	14.10	8.02	2.7
✓	✓	-	84.1 ± 0.5	14.2 ± 3.8	12.40	6.23	2.1
✓	-	✓	83.9 ± 0.4	13.7 ± 1.7	13.65	7.05	2.3
✓	✓	✓	**84.3** ± 0.4	**13.3** ± **1.6**	**12.34**	**6.11**	**2.0**
SSA:			84.1 ± **0.2**	14.7 ± 2.3	65.75	29.55	2.7

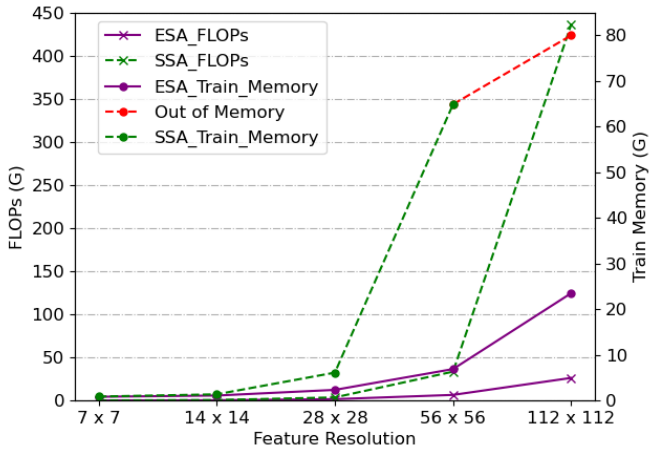


Figure 5: Comparison of computational complexity and memory cost between Efficient Self-Attention (ESA) and Standard Self-Attention (SSA) with gradually increased input feature resolution. The indicators are calculated from a Transformer branch with 5 layers and 8 heads.

We reproduce several newly proposed methods on Synapse or M&Ms based on their open-source codes and follow our training or evaluation strategies on the corresponding datasets. For all ablation experiments, we use the average accuracy obtained from training with five different random seeds to give more reliable conclusions.

5. Results

5.1. Comparison on the Synapse Dataset

We report comparison results with multiple previous state-of-the-arts methods on Synapse in Table 1. From these quantitative results, MultiTrans achieves leading performance on the DSC and HD with the third smallest model parameters. Although the models of UTNet and HiFormer-L are lighter than MultiTrans, the segmentation accuracy of our method is statistically significant higher than these two methods (statistically significant tests: Table 9). Compared to all other models except for the two mentioned above, we achieved higher accuracy with fewer parameters. In particular, our method outperforms the well-known TransUNet by 7.24% on DSC and 19.03 mm on HD with only 37.42% parameters.

Furthermore, MultiTrans achieves the highest segmentation accuracy on all organs. Fig. 4 shows the visualization comparisons of MulitTrans with several newly proposed Transformer-based methods. From small-scale (e.g. Aorta, Kidney) to large-scale organs (Stomach and Liver), our network has clearer boundaries and helps reduce intra- and inter-class confusion. Combined with the quantitative results in Table 1, this strongly demonstrates the effectiveness of our architecture for handling different organs with variable shapes and sizes.

Table 5: Comparison between different Efficient Self-Attention methods by adopting them on our proposed network architecture. All Transformers are built with the same number of heads, layers, and input feature channel dimensions to ensure a fair comparison. Low-Rank: refers to similar methods that adopt different downsampling operations to reduce the sequence length of Key and Value [35, 36, 37], and we use the one proposed in MissFormer [35] as a representation. DeLighT: The ESA proposed by TransHRNet [29]. FLOPs are calculated with an input resolution of 224 × 224.

Name	DSC ↑	HD ↓	TM (GB↓)	FLOPs (G↓)	Params (M↓)
Low-Rank [35]	83.9 ± 0.3	15.0 ± 2.1	25.44	24.40	47.1
DeLighT [29]	83.8 ± 0.4	16.3 ± **0.0**	32.61	23.07	44.2
Axial [38, 32]	84.0 ± 0.4	15.0 ± 1.9	15.83	25.14	47.1
Deformable [30, 34]	84.0 ± 0.3	14.4 ± 1.2	12.69	17.34	39.7
MultiTrans	**84.3** ± 0.4	**13.3** ± 1.6	**12.34**	**17.14**	**39.4**
SSA [39]	84.1 ± **0.2**	14.7 ± 2.3	65.75	42.30	41.9

5.2. Comparison on the ACDC Dataset

The experimental results in Table 2 show that our method can also reach state-of-the-arts results on the ACDC dataset, and achieves the highest segmentation accuracy on all cardiac structures except the right ventricle. This demonstrate the superiority and generality of our network on different image modalities. In addition, we compared with several recently proposed 3D medical segmentation networks on this dataset. From the comparison in Table 2, the accuracy of 3D networks is generally higher than that of 2D networks because they can extract 3D spatial information. It is worth noting that our proposed 2D network structure with fewer parameters can still outperform these Transformer-based 3D networks.

5.3. Comparison on the M&Ms Dataset

We conducted experiments on this dataset mainly to measure the cross-vendor robustness of MultiTrans. Consistent with the results on Synapse and ACDC datasets, MultiTrans also achieves the highest accuracy in one of the experiments in Table 3, which was trained and tested with only vendor A. As for the other cross-vendor experiment, we trained all models with data from vendor A and B, and tested on vendor A, B, C, and D. From the results in Table 3, our network not only obtains the best performance on seen vendor A and B, but also on the unseen vendor C and D, demonstrating the effectiveness in handling vendor differences. We attribute this to the design of multi-branch Transformers on different levels of CNN, enabling MultiTrans to be focused on global and local information at multiple scales.

Table 6: Experimental results of using a single Transformer branch on each individual level feature of CNN and the results of removing one of the four branches. Branch 1 to 4: refer to the Transformer branches built on the highest resolution L_1 (the lowest-level) to the lowest resolution L_4 (the highest-level) in Fig. 1(a) respectively. The performance degradation compared to our final model is shown in parentheses.

Operation	DSC ↑	HD ↓
None	84.3 ± 0.4	13.3 ± 1.6
Single Branch:		
Branch 1	**83.5** ± 0.6 (0.8↓)	**19.6** ± 3.7 (6.3↑)
Branch 2	80.8 ± **0.2** (3.5↓)	20.4 ± 1.2 (7.1↑)
Branch 3	68.9 ± **0.2** (15.4↓)	19.8 ± **1.0** (6.5↑)
Branch 4	58.6 ± 0.4 (25.7↓)	26.5 ± 3.8 (13.2↑)
Removed Branch:		
Branch 1	80.9 ± 0.7 (**3.4↓**)	15.3 ± 3.4 (2.0↑)
Branch 2	83.4 ± 0.7 (0.9↓)	16.6 ± 1.7 (**3.3↑**)
Branch 3	84.0 ± 0.3 (0.3↓)	15.6 ± 1.2 (2.3↑)
Branch 4	83.9 ± 0.3 (0.4↓)	15.2 ± 1.2 (1.9↑)

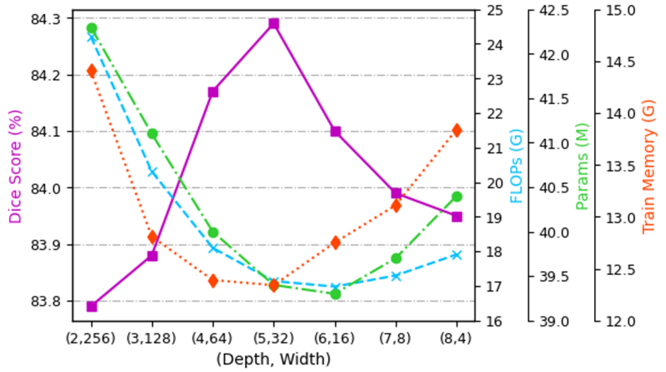


Figure 6: Comparison of accuracy and model complexity under different hyper-parameter choice of our Transformer branches. Depth: The number of layers in a Transformer branch. Width: The channel dimension of the Key-value in Self-attenion.

5.4. Ablation Studies

5.4.1. Ablation experiments with Efficient self-attention

Table 4 gives the ablation experiments on the design of efficient Self-Attention. It shows that the combination of Order-Changing (OC), Head-Sharing (HS) and Projection-Sharing (PS) can largely reduce the computation and training memory while maintain the accuracy or even slightly improve the accuracy. We attribute the accuracy improvement of HS and PS to the reduction of redundant features. Specifically, the training memory cost, FLOPs and Params of our Efficient Self-Attention (ESA) are 18.77%, 20.68% and 74.07% of the Standard Self-Attention (SSA).

Fig. 5 compared the computational complexity and memory cost between ESA and SSA under different input feature resolutions. With the increasing of spatial resolution, the training memory and computational complexity of the Transformer branch built with SSA exhibit a sharp exponential growth (the green and red lines), while our ESA has a moderate and approximately linear increase (the purple lines). Furthermore,

Table 7: Ablation experiments on design details in local-global feature fusion module and self-attention module. LS: long-skip connections. PE: sinusoid position embedding

Detailed Design	DSC ↑	HD ↓
Attention-Gated	84.3 ± 0.4	13.3 ± 1.6
w/o LS	83.7 ± 0.3 (0.6↓)	16.4 ± 1.2 (3.1↑)
Sum Fusion	83.9 ± 0.4 (0.4↓)	14.9 ± 1.1 (1.6↑)
w/o PE	84.1 ± 0.2 (0.2↓)	14.9 ± 2.3 (1.6↑)
w Scaling Factor	84.2 ± 0.2 (0.1↓)	15.2 ± 1.0 (1.9↑)

Table 8: Ablation experiments on the position of the Top-down path and the In-deep Supervision (IDS). The illustrations are given in Fig. 3.

Top-down Path	Supervision		DSC ↑	HD ↓
	Deep	IDS		
w/o	-	-	82.9 ± 0.5	22.1 ± 5.0
	✓	-	83.6 ± 0.3	16.7 ± 2.7
	-	✓	82.5 ± 0.4	21.8 ± 3.5
UNet-like	-	-	82.6 ± 0.4	20.5 ± 1.9
	✓	-	82.9 ± 0.6	15.7 ± 1.7
	-	✓	82.3 ± 0.5	17.8 ± 2.0
In-Deep	-	-	82.5 ± 0.4	23.8 ± 1.9
	✓	-	83.9 ± 0.3	14.9 ± 1.9
	-	✓	**84.3** ± 0.4	**13.3** ± **1.6**

when the input resolution increases from 28 × 28 to 56 × 56, the gap between ESA and SSA will significantly widen, with the difference of FLOPs increasing from 2.23× to 5.11× and the difference of memory cost increasing from 2.62× to 9.4×. This is why most existing methods choose to build Transformer branch on low-resolution features or adopt the tokenized image patches to reduce memory consumption and computation cost. Overall, our ESA can provide a more flexible design space for Transformer-based architectures, without being limited to building Transformer branches upon top-level feature maps of CNN or employing the tokenization technology.

We also compare several representative ESA methods widely used in medical image segmentation with our proposed ESA by adopting them on our proposed network architecture. From the results in Table 5, our proposed ESA achieves the highest accuracy with the lowest model complexity (training memory, computational complexity and Parameters) among all compared methods, indicating that it can be an effective and efficient alternative to the SSA or existing ESA methods. We attribute the advance of our ESA to two main reasons: (1): According to the three main steps of SSA (affinity matrix calculation, linear projections and value transformation), we propose three operations to reduce their costs respectively, while other methods only focus on saving the cost of computing the affinity matrix. (2) When computing the affinity matrix, we change the matrix multiplication order to reduce the complexity from quadratic to linear of the sequence length, which is mathematically equivalent to SSA. This operation avoids the information loss caused by the downsampling or point selection operations in other ESA

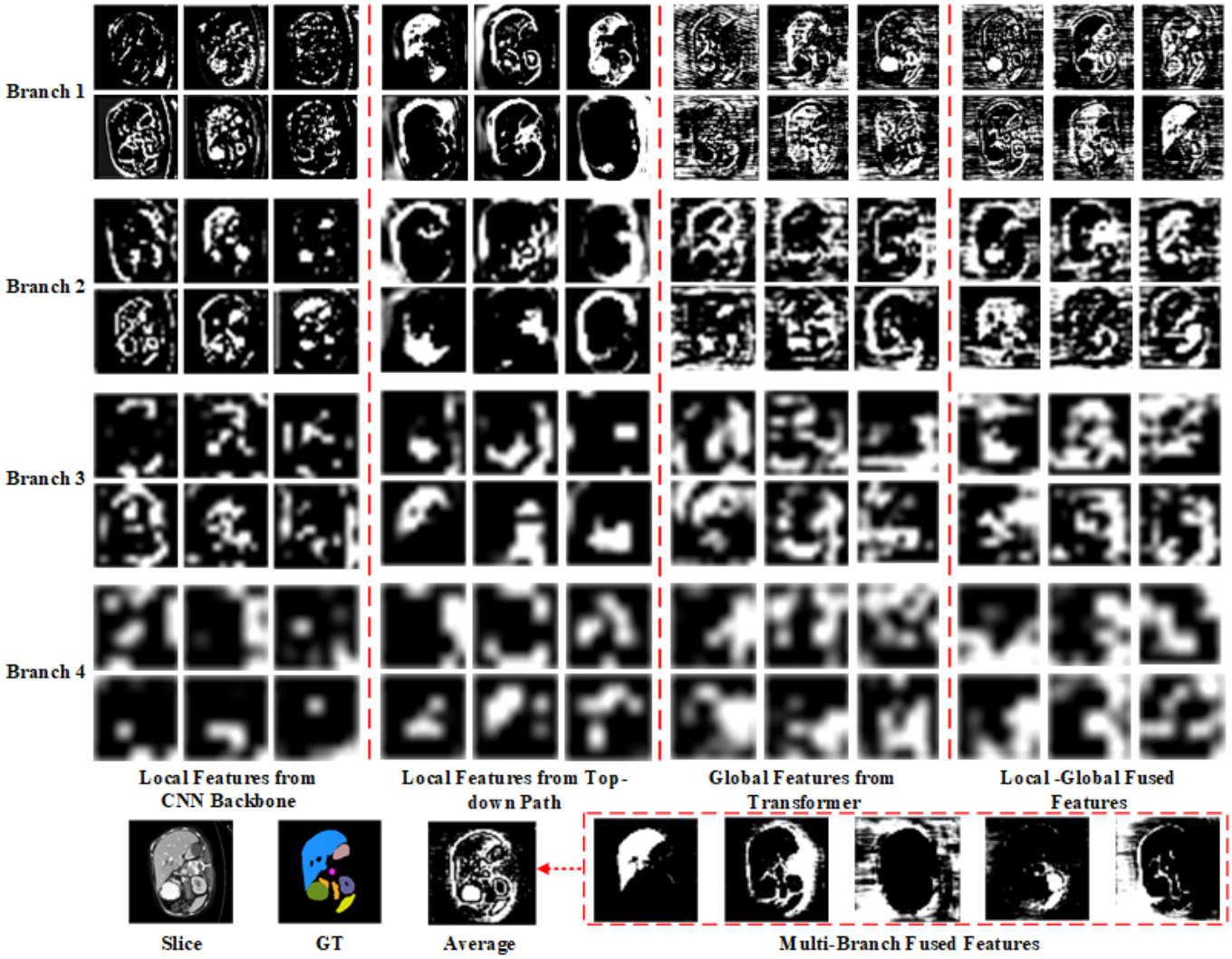


Figure 7: Visual analysis of the proposed network on Synapse dataset. We save feature maps from the critical stages of the trained network during inference, and extract 5 or 6 channel feature maps from hundreds of channels at equal intervals for visualization. 'Average': the average of **all channels** of the multi-branch fused features (the feature maps from the final stage of the network). Branch 1 to 4: refer to the Transformer branches built on the highest resolution L_1 (the lowest-level) to the lowest resolution L_4 (the highest-level) in Fig. 1(a) respectively.

methods, thus helping our ESA maintain competitive performance with SSA.

5.4.2. Ablation on the Multi-branch and Detailed Designs

Firstly, in Table 6, we use a single Transformer branch to perform inference on each individual level feature of the CNN, and Branch 1 to Branch 4 represent feature maps from the highest-resolution L_1 to the lowest-resolution L_4. Experimental results show that the segmentation performance of the Transformer branch gradually improves with the increasing of input feature resolution. This demonstrates the validity of our motivation that reasoning on relatively large resolution feature maps can improve segmentation accuracy by avoiding the loss of detailed information. Secondly, Table 6 also shows experimental results of removing one of the four branches. We can see that discarding the branch with the highest resolution results in a maximum decrease in segmentation accuracy (3.4%), consisting with our observations in the first experiment in Table 6. Besides, removing any of the four branches bring a minimum reduction of 0.3% for DSC and 1.9mm for HD, respectively, proving that each branch is complementary. We attribute the effectiveness of our multi-branch design to the fact that each level of CNN has a specific receptive field, which is more suitable for the segmentation of objects of a specific size. For example, the low-level features of CNN have a small receptive field required by small-scale objects, while the high-level features have a large receptive field required by large-scale objects.

Table 7 shows the effect of detailed design in local-global feature fusion module and self-attention module. By using long-skip connections (LS) to fuse local-global features, a simple 'Sum Fusion' operation slightly improves DSC and HD by 0.2% and 1.5 mm, respectively. We believe that the noise contained in different levels of CNN features (especially low-level features) reduces the effectiveness of local-global feature fusion operation [20]. After adding Attention-Gated to filter noise in

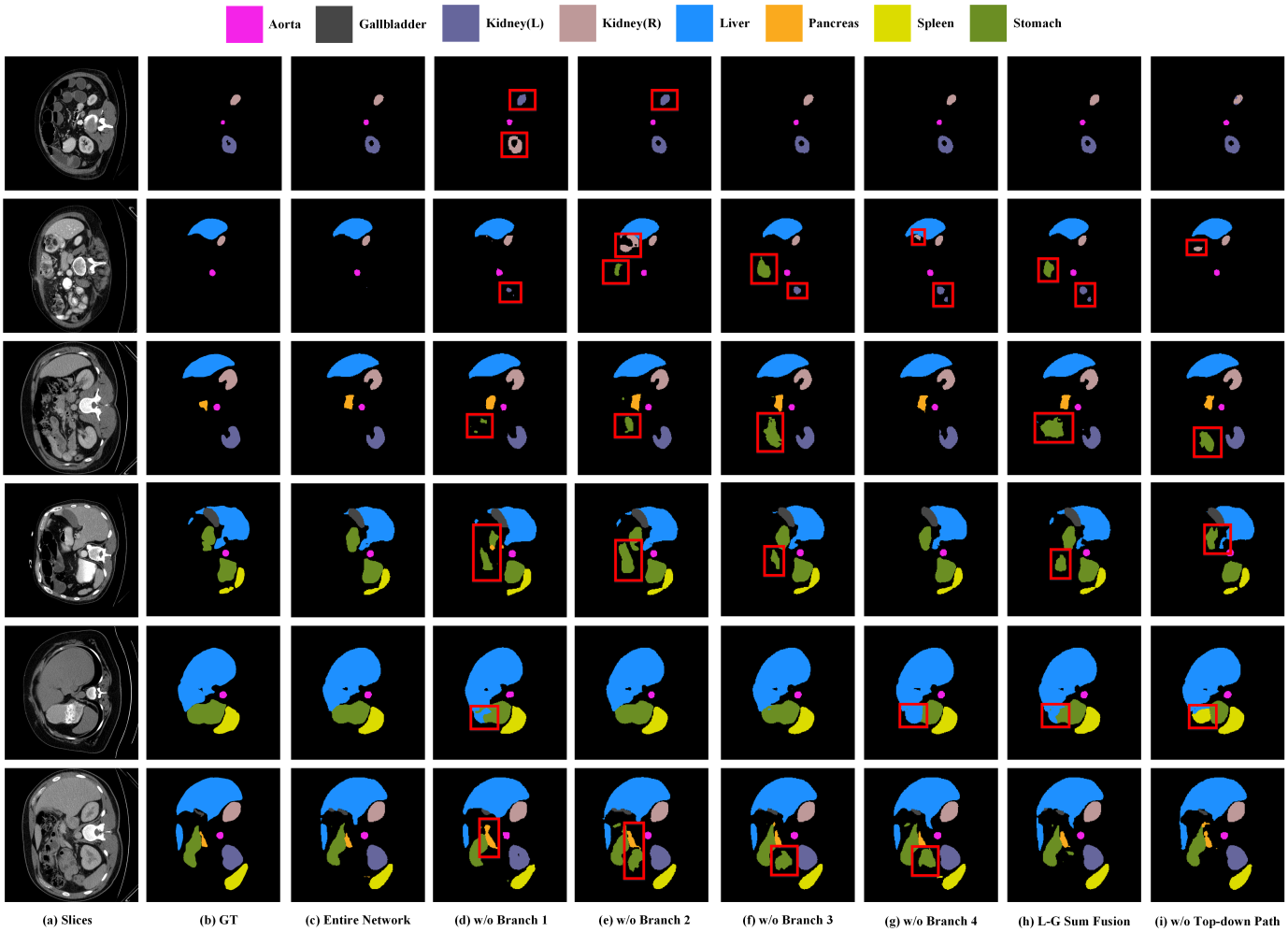


Figure 8: Qualitative results of the ablation studies on Synapse dataset. Branch 1 to 4: refer to the Transformer branches built on the highest resolution L_1 (the lowest-level) to the lowest resolution L_4 (the highest-level) in Fig. 1(a) respectively. L-G Sum Fusion: uses feature sum operation to replace the local-global feature fusion module based on Attention-gated. Red boxes mark the false positive or false negative predictions after removing the corresponding critical architecture of our MultiTrans.

local features, compared with without LS, DSC and HD are improved by 0.6% and 3.1 mm respectively. As for the detailed design of the self-attention module, the accuracy of DSC and HD dropped slightly by 0.2% and 1.6 mm after removing the position embedding, and we found that the Attention Scale operation used by SSA had a negative impact on the performance of our ESA.

Fig. 6 compares the accuracy and model complexity under different hyperparameter choice (deeper or wider) of our Transformer branches. Apparently, the 5-layer Transformer branch with Key-value of 32 channels achieves the best balance between performance and model complexity.

Table 9: Statistically significant test and model complexity comparison on the Synapse dataset. '$*$': The case-level DSC of the corresponding methods is provided by our replication. FLOPs are calculated with an input resolution of 224×224. The significant results (P-value < 0.05) are bolded.

Name	Case-level DSC	Params	FLOPs	P-value
	Mean ± SD	(M)	(G)	
LeViT-UNet-384* [48]	76.57 ± 9.03	52.2	25.55	**6.56E-04**
TransUNet* [21]	77.87 ± 9.38	105.3	29.30	**9.63E-03**
UTNet* [37]	78.11 ± 9.45	**9.5**	13.52	**3.69E-03**
TransFuse-L* [20]	80.48 ± 8.70	143.6	62.04	**1.74E-02**
MISSFormer [35]	81.96 ± 7.93	42.5	**9.89**	5.38E-02
MultiTrans	**84.82** ± **7.48**	39.4	17.14	-

5.4.3. Ablation on the Top-down path and the Deep Supervision

The comparison results of top-down paths and the deep supervision at different locations are shown in Table 8. The illustrations of them are given in Fig. 3. Here, UNet-like design refers to building the top-down path upon the outputs of multiple Transformer branches, while "In-Deep" design means adding the top-down path at different levels of the CNN backbone, in front of the Transformer branches. The segmentation accuracy of the UNet-like design is lower than our initial design (w/o top-down path, Fig. 3(a)) in all three experimental scenarios with different types of deep supervision. This demonstrates that the top-down path has a negative effect on the fusion of multi-scale global features, which we attribute to the convolu-

tional layers of the top-down path blurring the global features obtained by the Transformer. Furthermore, from the first and third rows of results for our initial design and UNet-like architecture, the In-deep Supervision degrades segmentation performance. The reason could be that the In-deep Supervision of these two designs directly adds additional supervision to different level features of the CNN backbone, while the low-level features are too shallow to learn semantic information.

Based on the above two observations, as shown in Fig. 3(c), we use a top-down path upon the CNN backbone to flow the topmost semantic features to low-level features. This design has two advantages: firstly, the top-down path helps deepen the backpropagation path of low-level features, benefitting In-deep Supervision; Secondly, it integrates multi-level features of the backbone to provide hierarchical features for low-level Transformer branches. In addition, we directly use large-scale bilinear upsampling to fuse features from Transformer branches to avoid the obtained global features being blurred by convolutional layers. Combined with In-deep Supervision, our final design achieved the highest accuracy on both DSC and HD in Table 8.

5.4.4. Visual analysis of the proposed network architecture.

To validate the rationality of the network architecture, we present feature maps from the critical stages of the trained network during inference in Fig. 7, from which we have the following observations: (1) The convolutional features of CNN or the Top-down path are focused in local and small regions, while the features from Transformers are activated throughout the entire field of view. This indicates that these two kinds of representations are complementary, and the Local-Global fused feature maps also show that the global features are enhanced on the details and boundaries, especially on Branch 1. (2) The comparison between local features from the CNN backbone and the Top-down path demonstrates that this design helps provide high-level semantics for low-level features (Branch 1 and 2). (3) It is clear that the local and global feature maps on different branches focus on the different regions, which proves that the proposed multi-branch Transformer architecture can provide both multi-scale global and local clues for the final prediction.

Furthermore, in Fig. 8, the qualitative results of the ablation studies show that the entire network exhibits robustness to scale variations, and the visualization results are consistent with the quantitative results in Table 6, 7 and 8.

5.5. Model Complexity Comparison and Statistically Significant Tests

In this section, we compare the complexity of the proposed model with other methods in terms of parameters and FLOPs, and FLOPs is calculated based on the input of 224×224. For statistically significant tests, we reimplemented several methods to obtain the DSC for each case in the Synapse test dataset. Then, we use paired samples t-test to calculate the p-value of comparing the DSC of other methods with that of ours. From the results in Table 9, our model obtains the highest average DSC and the smallest SD among all methods, and the model complexity of MultiTrans is lighter than most methods except for UTNet and MISSFormer. Specifically, our model is statistically significant higher than TransFuse-L with 27.44% parameters and 27.54% FLOPs. As for UTNet, our MutliTrans is statistically significant higher than it on DSC and has a smaller standard deviation, proving our method is more effective and robust than it. Although MISSFormer achieved the second best performance on Synapse (Table 1), out model still outperform it by a large margin on DSC (2.86%) with fewer parameters, and has a p-value of 0.0538, which is slightly higher than 0.05. Overall, our model achieves the best trade-off between performance and model complexity in terms of parameters and FLOPs.

6. Discussion

In this section, combined with the experimental results, we mainly discuss and summarize the overall performance of MultiTrans, the motivation and substantiation of the parallel Transformer branches design and the ESA module. As for the discussion and substantiation of the Top-down path design and the In-deep Supervision, we provide it in Section. 5.4.3.

The results in Table 1 and Table 2 show that our method can achieve state-of-the-arts accuracy compared to 2D or 3D segmentation networks on Synapse and ACDC datasets, demonstrating the superiority and generality of our network on different image modalities. The model complexity comparison and statistically significant tests in Table 9 further prove that our model achieves the best trade-off between performance and model complexity. Furthermore, from the results in Table 3, our network not only obtains the best performance on seen vendor A and B, but also on the unseen vendor C and D, showing strong cross-vendor robustness. We attribute this to the design of building parallel Transformer branches on different levels of CNN, enabling MultiTrans to learn both global and local features at multiple scales.

The purpose of designing the parallel Transformer branches is to handle the high shape and size variation of objects in medical images. Since different organs have different shapes and sizes, we prove that our design can achieve this purpose by providing the quantitative and qualitative comparison on the segmentation performance of various organs. In Table 1, MultiTrans achieves the highest segmentation accuracy on all organs and greatly improves the performance on low-precision organs (gallbladder and pancreas). Fig. 4 shows that, from small-scale (e.g. Aorta, Kidney) to large-scale organs (Stomach and Liver), our network clearly segments boundaries and reduces intra- and inter-class confusion. Combined with the quantitative results in Table 1, this strongly demonstrates the effectiveness of our architecture for handling different organs with variable shapes and sizes.

As for proposing the ESA module, our motivation is to reduce the computational complexity and train memory cost so that the Transformer branch can reason on large spatial resolution features to reduce the loss of fine spatial details and thereby improve segmentation accuracy. Fig. 5 explains that why most

existing methods have to build Transformer on low-resolution features or adopt the tokenized image patches to reduce memory consumption and computation complexity at the cost of sacrificing accuracy. When the input feature resolution increases from 28×28 to 56×56, the memory cost of the SSA-built Transformer rises sharply from 6.12 GB to 64.89 GB, exceeding the available memory of most GPUs. With the increasing of spatial resolution, the training memory and computational complexity of the Transformer built with SSA exhibit a exponential growth, while our ESA has a moderate and approximately linear increase. From Table 4, the MultiTrans built by our proposed ESA can largely reduce the training memory and computational complexity while even slightly improve the accuracy compared to the same architecture built by SSA. In addition, the experimental results in Table 6 indicate that as the input feature resolution increases, the single Transformer branch gradually improves the segmentation performance of the entire network. This proves that inference on relatively high-resolution feature maps indeed improves segmentation accuracy. Overall, our ESA provides a more flexible design space for Transformer-based networks, without being limited to building Transformer branches upon low-resolution feature maps or employing the tokenization technology.

Our network has some limitations that can be improved. Currently, we use four identical Transformer branches with the same number of layers and channel dimensions on different levels of backbone, which has achieved a good balance between accuracy and model complexity. However, from the 'Wider or Deeper' design guidelines of CNN [62, 63], our multi-branch Transformer network has the potential to achieve a better trade-off by choosing different depths and feature dimensions for different branches. In addition, the purpose of proposing Head- and Projection-Sharing is to reduce the cost of value transformation and linear projection operations in SSA, respectively. Interestingly, we find from the experiments in Table 4 that the combination of these two operations can also slightly improve the accuracy, indicating that the three linear projections and the multi-head affinity matrix may be redundant operations in SSA. We believe that this observation deserves further study as it could help us better understand the mechanisms of SSA and revisit its design. Another limitation of the research is that our network is a 2D network, which cannot extract 3D spatial information. Therefore, one future work is to build a 3D MultiTrans to see if our architecture design and the proposed ESA can benefit 3D networks.

7. Conclusion

In this work, we propose a novel Multi-Branch Transformer architecture for medical image segmentation. By building four parallel Transformer branches on different levels of CNN, our hybrid network aggregates both multi-scale global contexts and multi-scale local clues to provide hierarchical features for final prediction. To fit this architecture into accessible GPUs, we design a memory- and computation-efficient self-attention module to make it feasible to handle large-resolution inputs efficiently. The experiments on Synapse multi-organ segmentation dataset show the prediction of MultiTrans can favor various organs.

Statements of Ethical Approval

We used three published datasets, so ethical approval was not required.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported in part by Politecnico di Torino and in part by Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University.

References

[1] Z. Liu, J. Hou, X. Pan, R. Zhang, Z. Shi, Pa-net: A phase attention network fusing venous and arterial phase features of ct images for liver tumor segmentation, Computer Methods and Programs in Biomedicine (2023) 107997.

[2] Z. Zhu, X. He, G. Qi, Y. Li, B. Cong, Y. Liu, Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal mri, Information Fusion 91 (2023) 376–387.

[3] Y. K. Tsehay, N. S. Lay, H. R. Roth, X. Wang, J. T. Kwak, B. I. Turkbey, P. A. Pinto, B. J. Wood, R. M. Summers, Convolutional neural network based deep-learning architecture for prostate cancer detection on multiparametric magnetic resonance images, in: Medical imaging 2017: Computer-aided diagnosis, Vol. 10134, SPIE, 2017, pp. 20–30.

[4] J. Cao, H. Lai, J. Zhang, J. Zhang, T. Xie, H. Wang, J. Bu, Q. Feng, M. Huang, 2d–3d cascade network for glioma segmentation in multisequence mri images using multiscale information, Computer Methods and Programs in Biomedicine 221 (2022) 106894.

[5] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, H. D. Johansen, Kvasir-seg: A segmented polyp dataset, in: MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26, Springer, 2020, pp. 451–462.

[6] X. Liu, L. Song, S. Liu, Y. Zhang, A review of deep-learning-based medical image segmentation methods, Sustainability 13 (3) (2021) 1224.

[7] X. He, G. Qi, Z. Zhu, Y. Li, B. Cong, L. Bai, Medical image segmentation method based on multi-feature interaction and fusion over cloud computing, Simulation Modelling Practice and Theory 126 (2023) 102769.

[8] L. Qian, C. Wen, Y. Li, Z. Hu, X. Zhou, X. Xia, S.-H. Kim, Multi-scale context unet-like network with redesigned skip connections for medical image segmentation, Computer Methods and Programs in Biomedicine 243 (2024) 107885.

[9] F. Milletari, N. Navab, S.-A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 fourth international conference on 3D vision (3DV), Ieee, 2016, pp. 565–571.

[10] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, 2015, pp. 234–241.

[11] R. Wang, S. Chen, C. Ji, J. Fan, Y. Li, Boundary-aware context neural network for medical image segmentation, Medical Image Analysis 78 (2022) 102395.

[12] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: Redesigning skip connections to exploit multiscale features in image segmentation, IEEE transactions on medical imaging 39 (6) (2019) 1856–1867.

[13] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, J. Wu, 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), IEEE, 2020, pp. 1055–1059.

[14] X. Xiao, S. Lian, Z. Luo, S. Li, Weighted res-unet for high-quality retina vessel segmentation, in: 2018 9th international conference on information technology in medicine and education (ITME), IEEE, 2018, pp. 327–331.

[15] Q. Huang, L. Zhao, G. Ren, X. Wang, C. Liu, W. Wang, Nag-net: Nested attention-guided learning for segmentation of carotid lumen-intima interface and media-adventitia interface, Computers in Biology and Medicine 156 (2023) 106718.

[16] Q. Huang, L. Jia, G. Ren, X. Wang, C. Liu, Extraction of vascular wall in carotid ultrasound via a novel boundary-delineation network, Engineering Applications of Artificial Intelligence 121 (2023) 106069.

[17] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, A. Dosovitskiy, Do vision transformers see like convolutional neural networks?, Advances in Neural Information Processing Systems 34 (2021) 12116–12128.

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

[19] Y. Li, Z. Wang, L. Yin, Z. Zhu, G. Qi, Y. Liu, X-net: a dual encoding–decoding method in medical image segmentation, The Visual Computer (2023) 1–11.

[20] Y. Zhang, H. Liu, Q. Hu, Transfuse: Fusing transformers and cnns 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, Springer, 2021, pp. 14–24.

[21] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, Y. Zhou, Transunet: Transformers make strong encoders for medical image segmentation, arXiv preprint arXiv:2102.04306 (2021).

[22] F. Yuan, Z. Zhang, Z. Fang, An effective cnn and transformer complementary network for medical image segmentation, Pattern Recognition 136 (2023) 109228.

[23] J. Wang, L. Wei, L. Wang, Q. Zhou, L. Zhu, J. Qin, Boundary-aware transformers for skin lesion 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, Springer, 2021, pp. 206–216.

[24] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, R. Girshick, Early convolutions help transformers see better, Advances in neural information processing systems 34 (2021) 30392–30400.

[25] W. Zhang, Z. Huang, G. Luo, T. Chen, X. Wang, W. Liu, G. Yu, C. Shen, Topformer: Token pyramid transformer for mobile semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12083–12093.

[26] Z. Zhu, M. Sun, G. Qi, Y. Li, X. Gao, Y. Liu, Sparse dynamic volume transunet with multi-level edge fusion for brain tumor segmentation, Computers in Biology and Medicine (2024) 108284.

[27] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, D. Zhang, Ds-transunet: Dual swin transformer u-net for medical image segmentation, IEEE Transactions on Instrumentation and Measurement 71 (2022) 1–15.

[28] S. Wang, B. Z. Li, M. Khabsa, H. Fang, H. Ma, Linformer: Self-attention with linear complexity, arXiv preprint arXiv:2006.04768 (2020).

[29] Q. Yan, S. Liu, S. Xu, C. Dong, Z. Li, J. Q. Shi, Y. Zhang, D. Dai, 3d medical image segmentation using parallel transformers, Pattern Recognition 138 (2023) 109432.

[30] Y. Xie, J. Zhang, C. Shen, Y. Xia, Cotr: Efficiently bridging cnn and transformer for 3d 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 III 24, Springer, 2021, pp. 171–180.

[31] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable detr: Deformable transformers for end-to-end object detection, arXiv preprint arXiv:2010.04159 (2020).

[32] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, V. M. Patel, 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, Springer, 2021, pp. 36–46.

[33] J. Ho, N. Kalchbrenner, D. Weissenborn, T. Salimans, Axial attention in multidimensional transformers, arXiv preprint arXiv:1912.12180 (2019).

[34] Y. Ji, R. Zhang, H. Wang, Z. Li, L. Wu, S. Zhang, P. Luo, Multi-compound transformer for accurate biomedical 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, Springer, 2021, pp. 326–336.

[35] X. Huang, Z. Deng, D. Li, X. Yuan, Y. Fu, Missformer: An effective transformer for 2d medical image segmentation, IEEE Transactions on Medical Imaging (2022).

[36] A. Shaker, M. Maaz, H. Rasheed, S. Khan, M.-H. Yang, F. S. Khan, Unetr++: delving into efficient and accurate 3d medical image segmentation, arXiv preprint arXiv:2212.04497 (2022).

[37] Y. Gao, M. Zhou, D. N. Metaxas, Utnet: a hybrid transformer architecture 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 III 24, Springer, 2021, pp. 61–71.

[38] H. Wang, S. Xie, L. Lin, Y. Iwamoto, X.-H. Han, Y.-W. Chen, R. Tong, Mixed transformer u-net for medical image segmentation, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 2390–2394.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[40] B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jégou, M. Douze, Levit: a vision transformer in convnet's clothing for faster inference, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 12259–12269.

[41] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, D. Rueckert, Attention gated networks: Learning to leverage salient regions in medical images, Medical image analysis 53 (2019) 197–207.

[42] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-unet: Unet-like pure transformer for medical image segmentation, in: European conference on computer vision, Springer, 2022, pp. 205–218.

[43] D. Karimi, S. D. Vasylechko, A. Gholipour, Convolution-free medical image segmentation using transformers, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24, Springer, 2021, pp. 78–88.

[44] Y. Wu, K. Liao, J. Chen, J. Wang, D. Z. Chen, H. Gao, J. Wu, D-former: A u-shaped dilated transformer for 3d medical image segmentation, Neural Computing and Applications 35 (2) (2023) 1931–1944.

[45] Z. Shen, M. Zhang, H. Zhao, S. Yi, H. Li, Efficient attention: Attention with linear complexities, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2021, pp. 3531–3539.

[46] P. Hu, F. Perazzi, F. C. Heilbron, O. Wang, Z. Lin, K. Saenko, S. Sclaroff, Real-time semantic segmentation with fast attention, IEEE Robotics and Automation Letters 6 (1) (2020) 263–270.

[47] S. Fu, Y. Lu, Y. Wang, Y. Zhou, W. Shen, E. Fishman, A. Yuille, Domain adaptive relational reasoning for 3d multi-organ segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23, Springer, 2020, pp. 656–666.

[48] G. Xu, X. Wu, X. Zhang, X. He, Levit-unet: Make faster encoders with transformer for medical image segmentation, arXiv preprint arXiv:2107.08623 (2021).

[49] M. Heidari, A. Kazerouni, M. Soltany, R. Azad, E. K. Aghdam, J. Cohen-Adad, D. Merhof, Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 6202–6212.

[50] A. He, K. Wang, T. Li, C. Du, S. Xia, H. Fu, H2former: An efficient hierarchical hybrid transformer for medical image segmentation, IEEE Transactions on Medical Imaging (2023).

[51] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and

pattern recognition, 2016, pp. 770–778.
[52] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 801–818.
[53] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2881–2890.
[54] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, L. Shao, Pranet: Parallel reverse attention network for polyp segmentation, in: International conference on medical image computing and computer-assisted intervention, Springer, 2020, pp. 263–273.
[55] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, K. H. Maier-Hein, nnu-net: a self-configuring method for deep learning-based biomedical image segmentation, Nature methods 18 (2) (2021) 203–211.
[56] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, D. Xu, Unetr: Transformers for 3d medical image segmentation, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2022, pp. 574–584.
[57] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, Y. Yu, nnformer: Interleaved transformer for volumetric segmentation, arXiv preprint arXiv:2109.03201 (2021).
[58] C. Yang, X. Guo, T. Wang, Y. Yang, N. Ji, D. Li, H. Lv, T. Ma, Automatic brain tumor segmentation method based on modified convolutional neural network, in: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2019, pp. 998–1001.
[59] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 3146–3154.
[60] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.
[61] V. M. Campello, P. Gkontra, C. Izquierdo, C. Martin-Isla, A. Sojoudi, P. M. Full, K. Maier-Hein, Y. Zhang, Z. He, J. Ma, et al., Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge, IEEE Transactions on Medical Imaging 40 (12) (2021) 3543–3554.
[62] Z. Wu, C. Shen, A. Van Den Hengel, Wider or deeper: Revisiting the resnet model for visual recognition, Pattern recognition 90 (2019) 119–133.
[63] W. Liu, A. Rabinovich, A. C. Berg, Parsenet: Looking wider to see better, arXiv preprint arXiv:1506.04579 (2015).