MISSU: 3D Medical Image Segmentation via Self-Distilling TransUNet

Nan Wang, Student Member, IEEE, Shaohui Lin, Member, IEEE, Xiaoxiao Li, Member, IEEE, Ke Li, Yunhang Shen, Yue Gao, Senior Member, IEEE, and Lizhuang Ma, Member, IEEE

Abstract—U-Nets have achieved tremendous success in medical image segmentation. Nevertheless, it may have limitations in global (long-range) contextual interactions and edge-detail preservation. In contrast, the Transformer module has an excellent ability to capture long-range dependencies by leveraging the self-attention mechanism into the encoder. Although the Transformer module was born to model the long-range dependency on the extracted feature maps, it still suffers high computational and spatial complexities in processing high-resolution 3D feature maps. This motivates us to design an efficient Transformer-based UNet model and study the feasibility of Transformer-based network architectures for medical image segmentation tasks. To this end, we propose to self-distill a Transformer-based UNet for medical image segmentation, which simultaneously learns global semantic information and local spatial-detailed features. Meanwhile, a local multi-scale fusion block is first proposed to refine fine-grained details from the skipped connections in the encoder by the main CNN stem through self-distillation, only computed during training and removed at inference with minimal overhead. Extensive experiments on BraTS 2019 and CHAOS datasets show that our MISSU achieves the best performance over previous state-of-the-art methods. Code and models are available at: https://github.com/wangn123/MISSU.git

Index Terms—Self-distillation, transformer, medical image segmentation, 3D convolutional neural networks.

I. INTRODUCTION

Medical image segmentation plays a vital role in computer-aided diagnosis [1], which has achieved remarkable success with the usage of convolutional neural networks (CNNs). Among various network architectures, UNet [2] has been a mainstream framework for medical image segmentation, as it uses simple yet effective skip-connections to leverage the low-level and high-level semantic features of the encoder into the decoder. Subsequently, substantial efforts have been devoted to further improving the performance of medical image segmentation by using variant UNets, (e.g., VNet3D [3], DAF [4], and nnUNet [5]), which have received ever-increasing research attention.

Due to the limited receptive field of UNets, it is difficult to establish an explicit long-distance dependency [6]. Inversely, the Transformer-based algorithm [7] makes the self-attention mechanism feasible on a global scale, which can effectively learn long-distance dependency. Consequently, researchers improved the feature representation by integrating UNet-based methods with Transformer modules. For example, Transformer-based UNets [6], [8] have been proposed to explicitly capture the long-distance dependency, expanding the limited receptive fields of convolutional kernels in vanilla UNet to learn global semantic information. Despite the merits they brought, these modules alone may only achieve sub-optimal performance due to the lack of consideration for a compensatory relationship between global and local feature representation. Especially, it is crucial to learn the global semantic information and detailed local features simultaneously for medical image segmentation. Significantly, the deep branches/layers should extract semantic context information from low-resolution inputs, while the shallow ones concentrate on capturing spatial details (e.g., texture and edge) from the high-resolution inputs. Multi-path CNNs ([9], [10]) directly fuse the path of global semantic features and the path of local spatial information to improve segmentation performance. However, the computational cost for multi-path inference is significantly high and there is an explicit trade-off between positioning accuracy and multi-path information. This arouses our rethinking: how to design a unified framework for segmentation that implicitly models global semantic contexts and local spatial-detailed information during training while being efficient at inference?
Inspired by the powerful empirical results of the Transformer-based models on visual tasks, their promising generalization and robustness characteristics, and their flexibility to model long-range interactions, we propose to self-distill a Transformer-based UNet for 3D Medical Image Segmentation (MISSU), which implicitly fuses the long-distance dependency in global semantic information and the multi-scale local spatial-detailed features and brings no cost during testing. Like UNet, the proposed MISSU is also built upon the encoder-decoder structure with skip connections, whose flowchart is presented in Fig. 1. Specifically, the encoder first employs 3D CNN to extract features while down-sampling the spatial size. As such, we can capture local 3D context information from the last 3D feature maps. In this way, rich local 3D context features are effectively embedded in feature representation. Then each of them is reshaped into a vector (i.e., token) and fed into Transformer layers to capture the long-distance dependency. Additionally, to well extract local information, we design a multi-scale fusion block that receives local 3D features and generates multi-scale fusion outputs. Moreover, the former fusion outputs are progressively aggregated with the current local 3D features to enhance feature presentation ability. Additionally, self-distillation is computation-free at inference.

Self-distillation is introduced to transfer the knowledge of multi-scale local fusion outputs to local 3D features, which helps local 3D features converge to the same layers. Self-distillation is formulated by constraining the difference between the local feature and the multi-scale fusion edge of multi-scale local fusion outputs to local 3D features.

As discussed in Sec. I, local-to-global feature modeling is essential for medical image segmentation. There are various works either enhancing local features or modeling global feature representation [11]. Local feature enhancement is often implemented by image/feature pyramids with multi-scale information fusion [12], which has also been applied to medical image segmentation [13]. For example, CPFNet [13] designs a scale-aware pyramid fusion (SAPF) module to dynamically fuse multi-scale context information in a top-down manner. MLCFC [14] employs a multi-scale 3D CNN to incorporate both local and larger contextual information. Different from the former, MISSU is computation-free at inference. Extensive experiments demonstrate the superior performance of the proposed MISSU framework for medical image segmentation. On the widely-used BraTS 2019 dataset, our method achieves 89.98%, 85.77% and 80.14% on the segmentation of whole tumor, tumor core and enhanced tumor, respectively, notably outperforming state-of-the-art methods.

The remainder of this paper is organized as follows: In Sec. II, related works about medical image segmentation, Transformer module and self-distillation learning are introduced. The key components of the proposed MISSU are described in Sec. III, such as the network encoder (Sec. III-B), local-to-global feature modeling by self-distillation (Sec. III-C), and network decoder (Sec. III-D). Elaborate experiments and analysis are conducted in Sec. IV. We give the discussion of the proposed MISSU in Sec. V. Finally, the algorithm summary of MISSU is presented in Sec. VI.

II. RELATED WORK

A. Medical Image Segmentation

As discussed in Sec. I, local-to-global feature modeling is essential for medical image segmentation. There are various works either enhancing local features or modeling global feature representation [11]. Local feature enhancement is often implemented by image/feature pyramids with multi-scale information fusion [12], which has also been applied to medical image segmentation [13]. For example, CPFNet [13] designs a scale-aware pyramid fusion (SAPF) module to dynamically fuse multi-scale context information in a top-down manner. MLCFC [14] employs a multi-scale 3D CNN to incorporate both local and larger contextual information.
these methods, we propose a multi-scale fusion block to progressively refine the detailed local features in a bottom-up manner, which can be removed to reduce the computation cost by self-distillation at inference. Similarly, the global feature modeling also improves segmentation performance, which is often implemented by global interaction attention [15], [16] and self-supervised learning [17]. Firstly, the attention mechanism [18] can capture the long-distance dependence in the feature map to learn global feature representation. The attention gate model [16] can be integrated into U-Net architectures to increase the model sensitivity and segmentation performance. Equally, the attention-oriented U-Net model [15] replaces convolutional layers with reside-density modules via attention mechanism. Consequently, they fail to comprehensively capture interactions and similarities between subjects. To tackle this issue, we employ the Transformer and multi-scale fusion strategies to get local-to-global interaction attention. Then, self-supervised medical image segmentation aims to learn the global feature representation by designing various pretext tasks (e.g., solving jigsaw puzzles [19], [20], rotation prediction [21] and context restoration [22], [23]). However, self-supervised methods require two-stage training (i.e., pre-training and fine-tuning), which are computationally intensive for training. In contrast, our MISSU trains the model from scratch in an end-to-end manner, which is simple yet efficient for training.

**B. Transformer For Medical Image**

Transformer, as a new attention-driven building block [7], [24] has been applied in medical image segmentation, which establishes long-range dependence to capture context information. For example, UNETR [25] utilizes the pure Transformer as the encoder to learn sequence representations of the input volume, which captures the global multi-scale information. TransUNet [26] employs the Transformer to process each 3D medical image in a slice-by-slice manner, which cannot well learn continuous information between slices. To better use temporal information, TransBTS [6] uses 3D CNNs to extract the context information and employs a Transformer to model global feature representation with long-range dependence. Differently, our MISSU simultaneously learns global semantic information and refines local features to generate spatial-detailed features, further improving the segmentation performance. Unlike TransBTS, with the larger embedding dimension and more multi-head self attention, our MISSU can significantly reduce the computation overhead and parameter memory by the reduction of embedding dimension and multi-head self attention, and can also learn better global semantic features.

Equally, multi-path fusion learning between local and global features [27], [28] has also been applied to medical image segmentation. For example, Vessel-Net [29] introduces multi-paths to preserve the rich and multi-scale deep features during the model optimization. D-MEM [30] proposes a deformable multi-path ensemble for both local and global features for automated cervical nuclei segmentation. However, multi-path fusion is computation-intensive, as all local and global paths need to be computed during inference. Differently, we employ implicit fusions between local and global features by self-distillation so that the computation can be removed at inference.

**C. Self-Distillation Learning**

Self-distillation explores the potential of knowledge distillation from a new perspective [31]. Similarly, self-distillation strategies can be summarized as extracting the attention map of the current layers and then transferring the knowledge to the previous layers [32]. It aims to improve the performance of a compact model by using its knowledge without a teacher network. Self-distillation methods [33], [34] are mostly applied to language modeling, image recognition and object detection. For example, Hou et al. [35] presented a self-attention distillation approach for lane detection by allowing a network to utilize the attention maps of its own layers as distillation targets for its following layers. To our best knowledge, it is unexploited for medical image segmentation, except KD-ResUNet++ [36]. However, KD-ResUNet++ uses past predictions about data from the model to soften the targets at the previous epoch. It has a large number of parameters and is overly dependent on the previous period model. Inspired by the previous work, we employ the self-distillation framework in medical image segmentation, and improve the method so that it can learn the student network by transferring the knowledge online from its auxiliary information.

**III. Method**

**A. Overall Pipeline Of MISSU**

As illustrated in Fig. 1, the proposed network architecture of MISSU is built on the encoder-decoder framework. In the encoder, an input MRI scans \( X \in \mathbb{R}^{C \times H \times W \times D} \) with \( C \) channels (modalities), \( H \times W \) spatial resolution and \( D \) depth dimensions (slices) first goes through a 3D CNN-based encoder to generate local feature maps, capturing the spatial and depth information. Then, Transformer layers are used to model global feature representation with long-distance dependency. To compensate for the feature with detailed local information (e.g., shape and border of organs), we introduce the multi-scale fusion block that receives local features and generates multi-scale fusion outputs. Furthermore, self-distillation is proposed to transfer the knowledge from multi-scale fusion outputs to local features at the same layers during training, which can be removed at inference to reduce the computation cost. Finally, the decoder, consisting of multiple upsampling steps, decodes hidden features containing global and local information, and progressively produces the full resolution segmentation map. More detailed network architecture of MISSU is provided in our released code.\(^1\)

**B. Network Encoder**

Since the computational complexity of the Transformer is polynomial with respect to the amount of tokens (i.e., sequence length), simply flattening the input image to sequence as Transformer input is unfeasible. For example, ViT [7] applies...
Transformer on computer vision, in which tokens are often constructed by splitting an image into patches and fed into self-attention blocks to model long-range feature dependency. Specifically, ViT divides an image into the fixed-size $16 \times 16$ patches, and then reshapes each patch into a token, decreasing the sequence length to $16 \times 16$. For 3D volumetric input data with spatial and depth dimensions, the naive tokenization method will not model the detailed local information [37] across spatial and depth dimensions for volumetric segmentation.

To address the above problem, we first apply $3 \times 3 \times 3$ convolution blocks with downsampling to gradually encode the input image into a low-resolution/high-level feature representation $A^t \in \mathbb{R}^{N \times \frac{H}{s} \times \frac{W}{s} \times \frac{D}{s}}$ with channels $N = 16 \times 2^{t-1}, s = 1, \ldots, 4$, where $s$ is the stage number. We can extract the local detailed features in $A^t$. Subsequently, the input $A^t$ is fed into Transformer to learn the long-range correlation across spatial and depth dimensions with the global receptive field.

Transformer Block: We first construct tokens at the local feature $A^t \in \mathbb{R}^{128 \times \frac{H}{s} \times \frac{W}{s} \times \frac{D}{s}}$. We reshape $A^t$ into a sequence of volumetric vector/token $a_p \in \mathbb{R}^{128}$ according to the total length $M$ of spatial and depth dimension $D$, where $p = 1, 2, \ldots, M$ (i.e., $\frac{H}{s} \times \frac{W}{s} \times \frac{D}{s}$). To ensure a comprehensive presentation of each volume in $a_p$, a linear projection with weight $E$ is used to increase the channel dimension $N$ to $d$ in the embedding space. Following ViT [7], we also introduce specific learnable position embedding $E_{pos}$, which is added into embedding tokens for retaining positional information. Different from ViT, we remove the class token for the segmentation task. Therefore, we can formulate the above computation as:

$$Z_0 = \left[ a_p^1E; a_p^2E; \ldots; a_p^M E \right] + E_{pos},$$  \hspace{1cm} (1)

where $E \in \mathbb{R}^{512 \times 128}$ and $E_{pos} \in \mathbb{R}^{M \times 512}$ are the patch embedding projection and the position embedding, respectively. $Z_0 \in \mathbb{R}^{M \times 512}$ denotes the output feature embedding.

We then feed the feature embedding outputs $Z_0$ into the Transformer encoder, which consists of $L$ Transformer layers. A Transformer layer comprises a multi-head self-attention (MSA) block, and a feed-forward network (FFN). The output $Z_l$ in the $l$-th Transformer layer is defined as follows:

$$Z'_l = MSA \left( LN \left( Z_{l-1} \right) \right) + Z_{l-1}, \hspace{1cm} l = 1, 2, \ldots, L,$$

$$Z_l = FFN \left( LN \left( Z'_l \right) \right) + Z'_l, \hspace{1cm} l = 1, 2, \ldots, L,$$  \hspace{1cm} (2)

where $LN(\cdot)$ is the layer normalization.

Finally, the feature mapping consists of one linear project layer and reshape operation, which is applied to the Transformer output $Z_L$ to generate the global features $Z^A$ with the same size as local feature $A^t$.

C. Local-to-Global Feature Modeling By Self-Distillation

To successfully segment different lesion areas with small and significant shapes, the local features $A^t$ should be enhanced to extract multi-scale local detailed information. Therefore, we propose multi-scale fusion block (MSF) and self-distillation to obtain multi-scale preservation of local detailed features for computation-free inference.

1) Multi-Scale Fusion Block: Inspired by deeplabv3 [38], we propose a novel multi-scale fusion block (MSF), which is leveraged as a feature pyramid in a bottom-up manner. As shown in Fig. 2, the proposed MSF is different from the Atrous Spatial Pyramid Pooling (ASPP) module [38] based on two aspects: On the one hand, we add our MSF to the shallow and intermediate features rather than features from the top layer, which makes the local features learn multi-scale detailed information for segmentation; On the other hand, the pooling layer in ASPP is removed to reduce the information loss of detailed features. We also introduce a feature pyramid to construct the multi-level fused features at the bottom-up pathway rather than the top-down one in FPN [39].

Input local features $A^t, s = 1, 2, 3$ are first fed into MSF to generate multi-scale refined local feature maps $\tilde{A}^t, s = 1, 2, 3$, which utilizes four parallel branches $f^t_p$ with parameters $\theta^t_p (p = 1, \ldots, 4)$. Each branch is implemented by a serial of atrous convolutions to achieve different receptive fields of 3, 7, 9 and 19. In particular, the first branch uses one $3 \times 3 \times 3$ atrous convolution with an atrous rate of 1 to generate feature maps with the receptive field of 3. The second branch uses one $3 \times 3 \times 3$ atrous convolution with an atrous rate of 1 following by one $1 \times 1 \times 1$ regular convolution, which can generate feature maps with the receptive field of 7. Compared to the second branch, the third branch increases one more $3 \times 3 \times 3$ atrous convolution with an atrous rate of 1, which obtains the receptive field of 9. The fourth branch also has one more $3 \times 3 \times 3$ atrous convolution to produce the receptive field of 19, compared to the third one. The total four branches’ outputs $f^t_p(A^t; \theta^t_p)$ and input local features $A^t$ are then added to generate $\tilde{A}^t$. Then, feature pyramids are used to generate multi-level MSF outputs in a bottom-up manner, which can be formulated as:

$$B^1 = \tilde{A}^1; \hspace{0.5cm} B^s = W^A_A \otimes \tilde{A}^t + W^A_B \otimes B^{s-1}, \hspace{0.5cm} s = 2, 3,$$  \hspace{1cm} (3)

where $B^s$ is the refined output of the local feature at the $s$-th stage. $W^A_A$ and $W^A_B$ are 3D $1 \times 1 \times 1$ and $3 \times 3 \times 3$ convolution kernels with strides of 1 and 2, respectively.

2) Self-Distillation: To further reduce the computation of multi-scale fusion blocks for inference, we propose a self-distillation mechanism to transfer the knowledge from the MSF outputs $B^s$ to its corresponding local features $A^s$ during training, and the computation for MSF outputs would
where attention mapping spatial and depth dimension performing on: be skipped. Inspired by [40], we construct attention maps over Input: The parameters \( \theta_e, \theta_p, \theta_d \).

1. Initialize \( t = 0 \) and parameters \( \theta_e, \theta_p, \theta_d \).
2. repeat
3. Forward Pass:
   Choose a mini-batch from \( D \), and conduct forward propagation and loss computation with \( \theta_e, \theta_p, \theta_d \) via Eq. 6.
4. Backward Pass:
   Compute the gradient of \( \nabla \theta_e, \nabla \theta_p \) and \( \nabla \theta_d \) by the deviation of Eq. 6 with respect to \( \theta_e, \theta_p, \theta_d \), respectively.
5. Update:
   Update \( \theta_e, \theta_p, \theta_d \) by Adam with the poly learning rate strategy.
6. \( t := t + 1 \).
7. until Convergence or \( t \) reaches maximum iterations \( T \).

IV. Experiment

A. Experimental Setups

1) Datasets: We experiment on two widely-used medical image segmentation benchmarks: BraTS 2019 dataset [42] for brain tumor segmentation and CHAOS dataset [43] for liver segmentation.

BraTS 2019 contains 259 high-grade glioblastomas (HGG) patients and 76 low-grade glioblastomas (LGG) patients for training, and 125 cases for validation. Each patient has four image modalities, including T1-weighted (T1), post-contrast T1-weighted (T1ce), T2-weighted (T2) and Fluid Attenuated Inversion Recovery (FLAIR). Each modality has been aligned into the same area, which has a volume of \( 240 \times 240 \times 155 \). However, intensity normalization is required for the input data to ensure that the grey values of each image have the same distribution. Within each modality, z-score normalization is used for the foreground with non-zero voxel values in the medical images and corresponding labels. There are four different ROIs/classes of brain tumors: whole tumor (WT), tumor core (TC), enhanced tumor (ET), and background without tumor. As shown in Fig. 3(a), WT contains peritumoral edema (green part), enhancing tumor (yellow part), and the necrotic and non-enhancing tumor core (red part); TC means the yellow and red region; ET stands for the red region; background without tumor is displayed in black.

CHAOS dataset contains 40 two-modality (i.e., T1 and T2) image sequences, where 20 image sequences are used for training and the remainder for testing without annotations. Each sequence has 30-50 slices with a resolution of \( 256 \times 256 \). Since testing labels are not provided, we follow MMLAO [44] by splitting the training set into subsets of 15 and 5 subjects for training and testing. As shown in Fig. 3(b), there are only two ROIs for liver segmentation: liver (white part) and background/other organs (black part).

2) Evaluation Metric: For the quantitative analysis of experimental results, we consider multiple performance measurements, including Dice-score, and Hausdorff distance, which are widely used as a criterion for medical image segmentation. Dice-score measures the overlap between two samples on the target area, which can be calculated as:

\[
\text{Dice}(P, T) = \frac{|P \cap T|}{|P| + |T|},
\]

where \( P \) and \( T \) indicate the predicted region and the ground truth region, respectively. The larger the Dice-score, the better the segmentation performance. We employ Dice-score to evaluate the segmentation performance for WT, TC and ET on BraTS 2019 and the liver region on CHAOS. Dice ranges from zero to one, and the higher score is better.

To calculate the distance between segmentation boundaries, the Hausdorff distance (HD) is utilized. The highest value of the shortest least square distance \( d(p, t) \) between all points...
Fig. 3. Visual comparison for brain tumor segmentation on BraTS 2019 and liver segmentation on CHAOS. (a) Brain tumor segmentation results are yellow, green and red areas representing enhancing tumor (ET), peritumoral edema (ED) and the necrotic and non-enhancing tumor core (NCR/NET), respectively. (b) Liver segmentation results, where white and black areas denote liver and background/other organs, respectively. 

$$HD(PL, GT) = \left\{ \sup_{p \in \partial p} \inf_{t \in \partial t} d(p, t), \sup_{t \in \partial t} \inf_{p \in \partial p} d(t, p) \right\}. \quad (8)$$

3) Implementation Details: We implement our approach in PyTorch 1.7.0 with 8 NVIDIA 2080Ti GPUs. The models are trained using Adam optimizer by setting $\beta_1 = 0.9, \beta_2 = 0.999$, and $\epsilon = 10^{-5}$. The learning rate is initialized by 0.0004, decaying by each iteration with power 0.9 using the poly learning rate strategy. The batch size and total epochs are set to 4 and 1000, respectively. The balance parameter $\lambda$ is empirically set to 0.3.

In the training phase, the image sequences in the BraTS are randomly cropped into patches from $240 \times 240 \times 155$ to $128 \times 128 \times 128$, while the padding procedure may be required to maintain a consistent number of slices within each training case. And the image sequences are augmented with random horizontal flips across the axial, coronal and sagittal planes by a probability of 0.5. The input channel (modality) $C$ is set to 4. Furthermore, random intensity shifts between $[-0.1, 0.1]$ and scale between $[0.9, 1.1]$. For CHAOS dataset, we randomly crop its sequences into patches from $256 \times 256 \times 30$ to $128 \times 128 \times 30$, and $C$ is set to 2. A number of data augmentation techniques are used during training: rotations, scaling and Gaussian noise, which remain consistent with nnUNet [5].

B. Quantitative and Qualitative Results

1) Comparison Methods: We compare MISSU with UNet-based methods, knowledge distilled method, self-supervised method and Transformer-based methods. We adopt the open-source codes of all comparison methods and use grid search to decide their best hyper-parameters. We list details of each comparison method below.

- **Unet-based methods** (e.g., UNet [2], 3D UNet [45], nnUnet [5], DAF [4] and VNet3D [3]): the architectures consist of a contracting path to collect context and a symmetric expanding path to locate precisely.
- **KD-Net** [46] utilizes knowledge distillation to transfer knowledge from a trained multi-modal network (teacher) to a mono-modal one (student).
- **Supervoxel** [47] generates images to inpaint by employing self-supervised learning-based masking instead of random masking, and also focuses on the region-of-interest for brain tumor segmentation.
- **Transformer-based methods** (UNETR [25], MeTrans [48], TransUNet [26], TransBTS [6], MSSA [49], TransClaw [8]). UNETR, MeTrans, TransUNet and TransBTS are described in Sec. II. MSSA uses a guided self-attention mechanism to capture contextual dependencies. TransClaw integrates Transformer layers in the encoding part to exploit multi-scale information. More compared architectures are presented in Sec. V-B.

2) Quantitative Comparison: To begin with, we evaluate the segmentation performance on BraTS 2019, which is summarized in Tab. I. Particularly, our method improves by 3.29% on average, compared to the vanilla UNets (i.e., 3D

| Method       | WT(%) | TC(%) | ET(%) |
|--------------|-------|-------|-------|
| 3D UNet[16][45] | 87.38 | 72.48 | 70.86 |
| nnUNet[21][5]  | 89.67 | 84.01 | 78.55 |
| KD-Net[20][46] | 78.52 | 82.70 | 72.89 |
| Supervoxel[20][47] | 85.11 | 79.34 | 72.15 |
| UNETR[21][25]  | 79.00 | 75.82 | 60.62 |
| MTrans[18][48]  | 87.33 | 74.39 | 63.19 |
| TransUNet[21][26] | 89.48 | 78.91 | 78.17 |
| TransBTS[21][6] | 88.89 | 81.41 | 78.36 |
| MISSU(Ours)    | 89.98 | 85.77 | 80.14 |

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Unet, nnUnet). This suggests that Unet-based approaches are effective. However, they have limitations in long-distance context fusion, which may lose detailed information, such as edge and texture. In addition, KD-Net [46] and Supervoxel [47], employ knowledge distillation and self-supervised learning for brain tumor segmentation, respectively. MISSU outperforms KD-Net by 3.07% on TC. Compared with Supervoxel, we obtain the result of 85.77%, which is higher than that (i.e., 80.94%) of the TC-Dice. To this end, MISSU also achieves better performance than the best Transformer competitor TransBTS 2. Specifically, compared to TransBTS, MISSU achieves an average improvement of 1.09%, 4.36% and 1.78%, in terms of WT, TC and ET, respectively. These results suggest that our proposed MISSU self-distilling Transformer design can achieve better feature learning performance than the SOTA TransBTS.

We further evaluate the performance of liver segmentation on the CHAOS dataset, which is summarized in Table II. Our MISSU always outperforms the backbones and Transformer-based methods, especially with a large gap in performance on baselines. Specifically, our method achieves 14.30% and 6.63% Dice-score gains over DAF [4] and VNet3D [3], whose backbones of the encoder are built by 3D CNNs. Compared to Transformer-based TransClaw [8], our method also achieves better performance in live segmentation with 2.47% Dice-score gains. These results confirm our claim that MISSU obtains satisfying prediction performance with local-to-global modeling by self-distillation.

3) Qualitative Comparison: We use the officially released codes for all comparison methods and apply the parameter settings recommendations in the associated literature to ensure that all comparison methods operate effectively on each dataset.

Fig. 3(a) shows visual comparisons of brain tumor segmentation with different methods. MISSU achieves the best visual segmentation results compared to other baselines (e.g., Supervoxel and TransBTS). In particular, the necrotic predicted non-enhancing tumor core (red area) is correctly classified at the exact position and range, as shown in the second row. Moreover, the proposed MISSU can perform better on the local details in the fourth row to enhance tumor edge segmentation and improve each voxel classification. Fig. 3(b) presents some liver visual segmentation results, where the liver is marked in white. Compared to VNet3D [3] and MSSA [49], our method achieves the most accurate segmentation for contour details. For example, the proposed MISSU can accurately segment the liver boundary in the last column, achieving almost the same results as the ground truth. The main reason could be that MISSU can help extract robust features by local-to-global fusing. Consequently, our experiments show promising results in tackling the issue of long-range dependencies by leveraging the self-attention mechanism and weight sharing.

4) Model Complexity Comparison. For fairness, both comparative and our methods adopt the same input-level setting (4 × 128 × 128 × 128) to evaluate the GFLOPs and parameters. The results are summarized in Table III. We find that TransUNet with the CNN-Transformer mixed backbone requires more parameters while significantly reducing the computation cost, compared to 3D Unet and nnUnet with the CNN-based backbone. The explanation is that 3D convolutions in the CNN-based backbones occupy with larger computation overhead, compared to the self-attention mechanism in Transformers. Compared to TransUNet, our method (in the last row) reduces 85.57M parameters (10.5M vs. 96.07M) while significantly improving the model performance of WT, TC and ET Dice-scores of 0.5%, 6.86% and 1.97% (in Tab. I), respectively. We further evaluate the real running time between our method and other SOTA methods. We can observe that the running time is relatively consistent with GFLOPs, and our method achieves the best trade-off between the segmentation performance and model complexity. The explanation is that our MISSU can effectively learn local-to-global features by implicitly fusing the global semantic information and multi-scale local spatial-detailed features via self-distillation.

C. Ablation Study

To verify the effectiveness of MISSU, we conduct ablation studies to analyze different elements, including the Transformer, MSF and self-distillation. The results are summarized in Table IV and V, where the base model is 3D-CNN Unet with all X. Both MSF (local) and MSF (MS-output) use multi-scale fusion blocks, while the former uses local features for skip connection and the latter uses the output by itself for skip connection. We also provide qualitative comparison results on the BraTS dataset to effectively demonstrate the effects of different modules, as shown in Fig. 6.

1) Effect of Transformer: We evaluate the effect of the Transformer (i.e., with vs. without Transformer). As shown
in Tab. IV, compared to the base model, the model after adding Transformer (i.e., the second row) achieves significant performance improvement, especially with the increase of 2.99% WT score on BraTS 2019 and the decrease of 4.5219mm Hausdorff distance on CHAOS dataset. As can be seen from Table IV, the combination of the Transformer, MSF (local) and self-distillation (i.e., the last row) achieves WT gains of over 2.27% and lower WT Hausdorff of about 1.5mm compared to the combination of MSF (local) and self-distillation (i.e., the second row). It demonstrates the advantages of using Transformer to model global interactions.

2) Effect of Multi-Scale Fusion Blocks: We further evaluate the effectiveness of MSF. Compared to the base model, MSF (local) can learn local detailed features to improve the segmentation performance. For example, adding MSF (local) (i.e., the third row in Tab. IV) achieves 11.86% Dice-score gains over the base model on CHAOS. We also found in Tab. V, with the Transformer block, the MSF (local) (i.e., the third row) significantly decreases the Hausdorff scores of about 2.06mm WT, 1.23mm TC and 2.24 mm ET on BraTS, compared to that of only Transformer in the second row of Tab. IV. This is due to the fact that MSF (local) has the ability to model irregular-shaped deformation of lesion regions. In Tab V, we also observe that Transformer + MSF (MS-output) achieves the best performance, compared to only Transformer + MSF (local). However, it requires the heaviest computation and memory cost for inference, as shown in Tab. III. Transformer + MSF (local) achieves the reduction of 64.5 GFLOPs and 6.88M parameters for four-modality and the number of 128 slices, compared to that of Transformer + MSF (MS-output).

3) Effect of Self-Distillation: Self-distillation is also an essential element to improve the segmentation performance without extra computation cost at inference. As can be seen from Table IV, the self-distillation (the last row) refines local features to learn more multi-level and detailed features, which achieves at least 1.14% (e.g., 83.35% TC vs. 82.21% in the third row) Dice-score improvement on BraTS 2019, as well as the decrease of Hausdorff distance by at least 1.21mm. Similar performance improvement equipped with the Transformer can also be found in Tab. V by using self-distillation. Explicit MSF (MS-output) in the third row achieves a slightly higher Dice-score than that in the last row. However, as shown in Tab. III, the Transformer+MSF (local)+self-distillation (SD) directly reduces 64.5 GFLOPs and 6.88M parameters, which achieves a better trade-off between Dice-score and computation/memory cost.

4) Sensitivity Analysis: We further conduct experiments to investigate the sensitivity of hyper-parameter \( \lambda \), Transformer scale and the number of skip-connections. We first split the original training set into a training subset and a validation subset, and then evaluate the sensitivity by cross-validation.

The sensitivity of \( \lambda \). We can observe that the best selection of \( \lambda \) is set to 0.3. As can be seen from Fig. 4, from 0.1 to 0.3, increasing \( \lambda \) is able to improve the segmentation performance. The performance slightly drops when \( \lambda \) is set from 0.4 to 1. Thus, we report the performance of the proposed MISSU by setting \( \lambda \) to 0.3, unless otherwise specified.

The sensitivity of Transformer scale. The scale of Transformer is primarily determined by two hyper-parameters: the feature embedding space \( d \) and the number of Transformer layers \( L \). To verify the impact of Transformer scale on segmentation performance, we conduct experiments to analyze its sensitivity. For this purpose, we set the range of the numbers of layers embedding space as \( \{1, 4, 8\} \) and \( \{384, 512, 768\} \), respectively. As shown in Tab. VI, the best layer number \( L \) and embedding dimension \( d \) are set to 4 and 512. We also find that increasing the embedding dimension may not always improve the performance yet brings extra computational cost. It also demonstrates that pursuing model depth (i.e., repetitive stacking of Transformer layers) is not always the best choice for architectural design.

The sensitivity of the skip-connection number. As previously noted, integrating UNet like skip-connections assists in the enhancement of finer segmentation details by recovering low-level spatial information. The purpose of this ablation is to test the impact of adding different numbers of skip-connections. By varying the number of skip-connections to 0/1/2/3, the segmentation performance in Dice on BraTS dataset is summarized in Fig. 5. We can see that increasing the number of skip-connections improves segmentation performance. Inserting skip-connections to all three intermediate layers.
 TABLE VI
THE SENSITIVITY OF TRANSFORMER SCALE BY CROSS VALIDATION, (e.g., THE NUMBER OF TRANSFORMER LAYERS (L) AND FEATURE EMBEDDING SPACE (d)). VALUES REPRESENT THE MEAN(±STD) BY RUNNING THE MODELS 5 TIMES

| Layers (L) | Embedding space (d) | BraTS | CHAOS |
|-----------|---------------------|-------|-------|
|           | Dice(%) ↑         | HD(mm) ↓ |       |
| WT        | TC                 | ET    | WT    | HD(mm) ↓ | ET    |
| 1         | 82.65(1.9)         | 77.70(1.1) | 74.67(1.4) | 9.8225(0.7) | 7.8883(0.8) | 7.8113(0.7) |
| 8         | 84.35(1.2)         | 79.34(1.5) | 76.08(1.3) | 9.0202(0.4) | 6.6861(0.9) | 6.4547(0.6) |
| 4         | 88.78(1.4)         | 79.92(0.9) | 79.19(1.6) | 6.2550(0.6) | 6.3433(0.5) | 4.2004(0.5) |
| 4         | 90.00(1.7)         | 85.83(1.3) | 79.93(1.8) | 5.9159(0.7) | 4.0696(0.5) | 3.9785(0.3) |
| 4         | 89.35(2.0)         | 82.70(1.0) | 80.94(1.0) | 6.1133(0.8) | 5.6847(0.5) | 3.7587(0.3) |

Fig. 4. The sensitivity of hyper-parameter λ on BraTS dataset.

Fig. 5. The sensitivity of the skip-connection number using Dice-score on BraTS.

upsampling steps yields the best results. Due to the recovery of low-level spatial detail information, significant performance is achieved for the critical WT, ET and TC (90.00%, 85.83%, 79.93%). These results support our initial hypothesis that UNet-like skip-connections should be incorporated into the Transformer design to enable learning of precise low-level information.

D. Visualization Analysis Of The Proposed Modules

1) Analysis on Visual Attention: We further employ gradient-weighted class activation mapping (Grad-CAM) [50] to see the localization attention map from the final convolution layer, which highlights the critical regions of the image. As shown in Fig. 6, the Transformer block tends to learn the global feature presentation focusing on the larger attention region, while MSF (local) focuses on the detailed feature presentation with the more prominent boundary area of brain tumor. We leverage MSF into Transformer learned by self-distillation, which significantly concentrates on the brain tumor area for accurate segmentation. It indicates that our method can learn global semantic information and local spatial-detailed features.

2) Visualization on the MSF and Self-Distillation: Our main innovations lie in the MSF block and self-distillation. We visualize the segmentation results to evaluate their effectiveness. As shown in Fig. 7, the proposed MISSU without MSF or self-distillation can accurately segment the edema regions (green part), but it can not effectively handle the detailed parts, especially on the necrotic core (red part). The explanation is that the detailed spatial features can not be preserved without MSF, which leads to unstable segmentation on the necrotic core. Adding MSF but without self-distillation can not obtain satisfactory segmentation on the boundary of different regions. In contrast, our MISSU with MSF and self-distillation achieves almost the same segmentation results as the ground-truth, which benefits from the effective local-to-global feature modeling.

3) Visual Interpretation of Transformer Attention Maps: We further visualize the attention maps from different Transformer layers to investigate the effectiveness of the Transformer, as shown in Fig. 8. Shallow Transformer layers (e.g., the 1st and 2nd Transformer layers) in our MISSU concentrate on the region with a relatively wide range and the area with high
score scatters around the target brain tumor. In the deeper layers (e.g., the 4th Transformer layer), the closer to the lesion area, the higher the attention score. Despite the highly irregular shape, our method can still effectively learn useful the feature for accurate medical image segmentation.

V. DISCUSSION

A. Difference From Other Transformer-Based Methods

To make a thorough comparison between our MISSU and other Transformer-based methods, we further give a complete analysis to demonstrate the powerful potential of our method. Although SOTA methods (e.g.,UNETR, MeTrans TransUNet and TransBTS) can achieve good performance via CNNs and Transformer, our MISSU has a completely different learning mechanism. For example, it is different from UNETR [25] in directly utilizing the pure Transformer as a backbone to model global and local features. We model local-to-global feature interaction in an inverse manner, i.e., first CNN and then transformer. Additionally, MeTrans [48] directly models 3D volumetric images in the form of a sequence of 2D image slices, fusing the local features from different layers as the input for the Transformer block, while we progressively refine the local features by the proposed MSF block and use the final local features as input to the transformer layers. Similarly, both TransUNet [26] and TransBTS [6] directly reuse the coarse local features from CNNs to the decoder by the skip connection, while our MISSU transfers the delicate local features using the MSF block and self-distillation into the decoder, which can obtain the multi-scale local detailed features to improve the performance. Therefore, the proposed local-to-global interaction learning framework in MISSU is novel, especially on 3D local feature enhancement and transfer for the computation-free inference. Note that we can employ efficient Transformer variants in our MISSU to reduce memory and compute complexity while preserving accuracy.

B. The Influence Of MSF

We make a comparison between MSF (local) and MSF (MS-output), aiming to evaluate the effects of implicit and explicit MSF. MSF (local) denotes the multi-scale fusion block adding after the original 3D features but only using the original 3D features for skip connection to the decoder during training, which means the original 3D features are implicitly refined via MSF. Note that MSF (local) does not compute the MSF module for inference. In contrast, MSF (MS-output) directly uses the outputs of MSF for skip connections during both training and testing, which explicitly uses the refined features and needs more computation for inference. For the comparison, the proposed self-distillation with MSF (local) can capture shape-aware local details, achieving the best trade-off between Dice-score and computation/memory cost, compared to only MSF (local) and MSF (MS-output). We speculate that the local features generated by MSF may improve the long-range feature modeling of the Transformer, resulting in better performance.

C. Limitation and Future Work

There are several possible limitations and future directions we will further explore to improve our method, although it has already achieved superior segmentation performance in our demonstrated experiments. Specifically, for broader applications, we should consider (1) Real-time segmentation. The proposed self-distillation can significantly decrease the inference time, but there is still a gap in real-time segmentation. This is due to the heterogeneous operators and parameter redundancy in the proposed MISSU model. In the future, we will explore a pure transformer architecture to model local-to-global features and study model compression methods to reduce the parameter redundancy; (2) Efficient training. The training of the proposed MISSU method relies on the pixel-wise label for each slide, which is time-consuming and labor-intensive. In the future, we will consider several effective self-supervised learning methods to reduce the annotation cost and improve the generalization ability of the proposed MISSU model; (3) Superior performance. Multi-view medical images often exist yet are unexploited in the proposed MISSU. We will leverage multi-view images and multi-modality ones into our method to further improve the medical image segmentation performance.

VI. Conclusion

In this paper, we propose a novel 3D medical image segmentation framework by self-distilling a Transformer-based UNet (termed MISSU), which simultaneously learns global semantic information and local spatial-detailed features. In particular, the Transformer block is used to model global semantic context information by adding after the local feature mapping from the last 3D CNN layer. Meanwhile, a local multi-scale fusion block is proposed to refine local features extracted by the main CNN stem through self-distillation, which is only computed during training and removed at inference. The global features embedded by the Transformer are upsampled and fused into the refined local features in the decoder to generate the pixel-level prediction progressively. In particular, we analyze the effectiveness of the Transformer, MSF and self-distillation. We have comprehensively evaluated the performance of MISSU on BraTS 2019 and CHAOS datasets, which demonstrates superior performance gains over the state-of-the-art methods. In the future, we will integrate both MRI and non-image (e.g., age and sex) information into our MISSU framework for tumor identification.

REFERENCES

[1] J. Soti, I. Valavanis, S. G. Mougiakakou, S. Golemati, A. Nikita, and K. S. Nikita, “Computer aided diagnosis based on medical image processing and artificial intelligence methods,” Nucl. Instrum. Methods Phys. Res. A, Accel. Spectrom. Detect. Assoc. Equip., vol. 569, no. 2, pp. 591–595, Dec. 2006.
