Med-Tuning: Parameter-Efficient Transfer Learning with Fine-Grained Feature Enhancement for Medical Volumetric Segmentation

Anonymous authors
Paper under double-blind review

Abstract

Deep learning-based medical volumetric segmentation methods either train the model from scratch or follow the standard "pre-training then fine-tuning" paradigm. Although fine-tuning a pre-trained model on downstream tasks can harness its representation power, the standard full fine-tuning is costly in terms of computation and memory footprint. In this paper, we present the study on parameter-efficient transfer learning for medical volumetric segmentation and propose a new framework named Med-Tuning based on intra-stage feature enhancement and inter-stage feature interaction. Additionally, aiming at exploiting the intrinsic global properties of Fourier Transform for parameter-efficient transfer learning, a new adapter block namely Med-Adapter with a well-designed Fourier Transform branch is proposed for effectively and efficiently modeling the crucial global context for medical volumetric segmentation. Given a large-scale pre-trained model on 2D natural images, our method can exploit both the crucial spatial multi-scale feature and volumetric correlations along slices for accurate segmentation. Extensive experiments on three benchmark datasets (including CT and MRI) show that our method can achieve better results than previous parameter-efficient transfer learning methods on segmentation tasks, with much less tuned parameter costs. Compared to full fine-tuning, our method reduces the fine-tuned model parameters by up to $4\times$, with even better segmentation performance.

1 Introduction

Medical image segmentation, which aims to delineate tumors and sub-regions of organs from biomedical images, is capable of assisting doctors in making accurate clinical diagnoses and treatment planning. It is vital to improve the accuracy and efficiency of medical volumetric segmentation, since the widely adopted medical modalities, including computed tomography (CT) (Nguyen et al., 2015) and magnetic resonance imaging (MRI) (Huo et al., 2017), are all composed of 3D volumes, and plenty of practical applications (e.g. tumor segmentation and anomaly detection) are based on the corresponding segmentation of these modalities. Deep neural networks have become a popular tool for this task, including architectures based on convolutional neural networks (CNNs) (Ronneberger et al., 2015; Milletari et al., 2016; Zhang et al., 2018; Myronenko, 2018; Çiçek et al., 2016; Zhou et al., 2018; Oktay et al., 2018; Isensee et al., 2021) and Transformers (Chen et al., 2021b; Valanarasu et al., 2021; Huang et al., 2021; Wu et al., 2022; Wang et al., 2021c; Li et al., 2022). With the continuous improvement of model performance, the number of model parameters and corresponding training costs have increased greatly, especially the Transformer-based models. Besides, due to the great challenges in realizing effective model training, these methods usually benefit from fine-tuning the models pre-trained on larger-scale datasets (e.g. ImageNet (Deng et al., 2009)), but still tune all the model parameters, which results in further training costs. Therefore, we are interested in the question: *Is there a way to pursue a balance between the model performance and fine-tuning parameter efficiency?*

In the community of natural image processing, the ***"pre-training then fine-tuning"*** paradigm has become standard practice to boost the model performance on downstream tasks. Conventional fine-tuning schemes include *full fine-tuning* and *head fine-tuning*, which optimize either the entire network or only the specific head (e.g. Linear (He et al., 2022) and Partial (Yosinski et al., 2014)). Full

fine-tuning usually achieves higher accuracy but also a higher training cost. Recent studies (Jia et al., 2022; Chen et al., 2022; Pan et al., 2022; Yu et al., 2022; Zhang et al., 2023a; Xu et al., 2023; Sung et al., 2022) on parameter-efficient transfer learning (PETL) aim to strike a balance in between.

In this paper, we explore the potential of PETL for medical volumetric segmentation. Unlike natural image datasets, the scale of the acquired medical datasets is generally small because of high annotation costs. As a result, there are many strong pre-trained 2D models on large-scale natural image datasets but such pre-trained models are lacking in the medical domain. **As image-level annotations are much more easily acquired than pixel-level ones, the majority of off-the-shelf pre-trained models (either through strong supervision or self-supervision) are classification-based architectures.** While the very recent visual foundation model Segment Anything Model (SAM) (Kirillov et al., 2023) highlights its potential in the medical domain, directly transferring its full segmentation model may result in a constrained architectural design, possibly not meeting accuracy standards for diverse downstream tasks. **In contrast, using classification pre-trained backbones offers more flexibility and convenience.** *To provide a thorough evaluation in this work, we also incorporate SAM as a backbone candidate in our following experiment section, further showcasing the powerful potential of our method.* Therefore, the goal of this study is to understand how to adeptly adapt strong pre-trained classification models from 2D natural images to medical volumetric segmentation tasks, rather than just pursuing top-tier performance through intricate designs.

As shown in Fig. 1, there are two-fold gaps between the pre-training source domain and downstream target domain that require considerations to achieve successful PETL: **(1)** the modality gap between 2D natural images and 3D medical volumes; **(2)** the task gap between the pre-training classification task and the downstream segmentation task. To narrow these gaps, we propose to build a PETL framework for medical volumetric segmentation based on pre-trained classification models on natural images with an efficient plug-and-play block to exploit the crucial spatial multi-scale features and volumetric correlations.
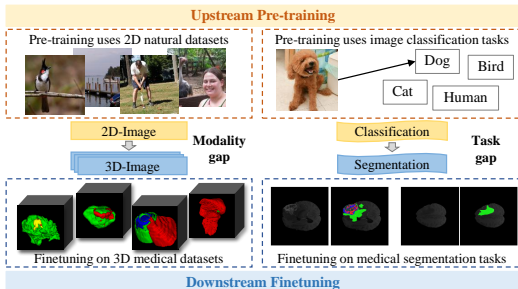

Figure 1: The illustration of the two-fold gaps between source and target domain when exploring pre-trained model on large-scale 2D natural image datasets for medical volumetric segmentation.

Specifically, for the first gap brought by 3D medical data itself, there is an essential temporal continuity that needs to be exploited between adjacent medical image slices. To address this, we design an adapter block (i.e. Med-Adapter) with high efficiency and flexibility while jointly conducting spatial-temporal (slice) modeling. For the second gap of taking semantic segmentation as the downstream task, previous studies (Long et al., 2015; Szegedy et al., 2015; Chen et al., 2014; 2017a;b; Wang et al., 2021a; Liu et al., 2021; Guo et al., 2022; Wang et al., 2021b) have shown that such dense prediction requires crucial multi-scale information, As a vital aspect of the multi-scale features, the global information counts a lot for the dense prediction tasks, while the Fast Fourier Transform (FFT) and Inverse Fast Fourier Transform (IFFT) naturally have a global vision due to their internal operation mechanism (more details can be found in Sec. 3.1), which is right on demand. Thus, by leveraging the intrinsic global vision characteristic of the FFT and IFFT, high-efficiency multi-scale branches coupled with the FFT branch (i.e. global branch) are effectively leveraged in our method for intra-stage feature enhancement and inter-stage feature interaction.

The main contributions can be summarized as follows:

- We present a study on PETL for medical volumetric segmentation and propose a new framework *Med-Tuning*, achieving the trade-off between segmentation accuracy and parameter efficiency.
- A new *Medical Adapter (**Med-Adapter**)* is proposed for PETL, as a plug-and-play component to simultaneously consider both multi-scale representations and inter-slice correlations.
- Our framework is generic and flexible, which can be easily integrated with common Transformer-based architectures to greatly reduce training costs and simultaneously boost model performance.
- Extensive experiments on three benchmark datasets with different modalities (e.g. **CT** and **MRI**) validate the effectiveness (e.g. Fig. 2) of our Med-Tuning over full fine-tuning and previous PETL methods for medical volumetric segmentation.

2 RELATED WORK

2.1 MEDICAL VOLUMETRIC SEGMENTATION

Unlike natural images, medical images pose unique challenges, such as uneven foreground-background distribution and sharp variations in lesion shape and scale. As proved by previous works (Çiçek et al., 2016; Milletari et al., 2016; Zhou et al., 2018; Chen et al., 2021a;b; Oktay et al., 2018), extracting multi-scale representations is crucial for desired segmentation performance. For example, Ronneberger et al. (Ronneberger et al., 2015) concatenated multi-scale features from CNN encoder and up-sampled features, complementing the loss of spatial information caused by down-samplings. Cao et al. (Cao et al., 2021) also used skip connec-

Figure 2: Comparison with previous PETL methods in terms of the trade-off between the number of tuned parameters and segmentation accuracy. The backbone ViT-B/16 is pre-trained on ImageNet-21k and fine-tuned on BraTS2019 dataset.

tions to gradually fuse the low-level features and the high-level features in Transformer architecture.

In addition, the information between continuous slices (i.e. volumetric correlation) of medical images is of critical importance. Various medical volumetric segmentation methods (Çiçek et al., 2016; Milletari et al., 2016; Zhou et al., 2018; Oktay et al., 2018) have effectively explored this vital continuity by utilizing 3D convolutions (Çiçek et al., 2016; Milletari et al., 2016; Zhou et al., 2018) or introducing self-attention mechanism among the 3D input patches (Wang et al., 2021c).

Based on the above analysis, the proposed framework Med-Tuning simultaneously takes both multi-scale feature representations and inter-slice correlation into consideration, realizing the effective spatial feature and temporal relationship modeling in a parameter-efficient manner by the simple yet effective PETL architecture.

2.2 PARAMETER-EFFICIENT TRANSFER LEARNING

Conventional fine-tuning methods fail to achieve the trade-off between accuracy and parameter efficiency. Therefore, various PETL methods were born on demand recently, which can be summarized into three categories: The *first* one is Prompting (Bahng et al., 2022), which modifies the input pixel space of Transformer layers. VPT (Jia et al., 2022) added learnable prompts to patch embeddings for downstream visual tasks. However, VPT's sensitivity to prompt number and token length may limit parameter efficiency for dense prediction tasks. Pro-tuning (Nie et al., 2022) inserted multiple stage-wise prompt blocks into different stages of the backbone. DePT (Zhang et al., 2023a) tackles the Base-New Tradeoff dilemma in prompt tuning from a feature decoupling perspective for achieving better zero-shot generalization on new tasks.

The *second* type is Adapter that can be easily inserted into backbones. Specifically, Adapt-Former (Chen et al., 2022) replaced the original MLP block in Transformer with the proposed AdaptMLP. Despite its promising results, AdaptFormer overlooks volumetric information, which may lead to the loss of connections between video clips. To tackle this problem, ST-Adapter (Pan et al., 2022) injected Adapter-like blocks in each Transformer layer and introduced the 3D depthwise convolution (Ye et al., 2018) to capture spatial-temporal features. However, it does not take multi-scale representation modeling into account, which is critical for the segmentation task. Aim (Yang et al., 2023) implements spatial, temporal, and joint adaptation to enhance spatiotemporal reasoning in image models.

The *third* category includes other PETL techniques, e.g., LoRA (Hu et al., 2021) inserted learnable low-rank matrices into the self-attention block in Transformer, while V-PETL (Yu et al., 2022) extended the parameters of prefix tuning (He et al., 2021) from randomly initialized to input associated. SAN (Xu et al., 2023) is a small network that uses shortcut connections to take intermediate activations from backbone networks and make predictions. TransSeg (Zhang et al., 2023b) employs a weight inflation strategy to transition pre-trained Transformers from a 2D to a 3D context, preserving the advantages of both transfer learning and depth of information.

Nevertheless, the above-mentioned research mainly focuses on the 2D/3D classification tasks on natural images. Few of these works make targeted structural improvements for downstream dense prediction tasks like segmentation. Besides, as analyzed above, none of the previous works has
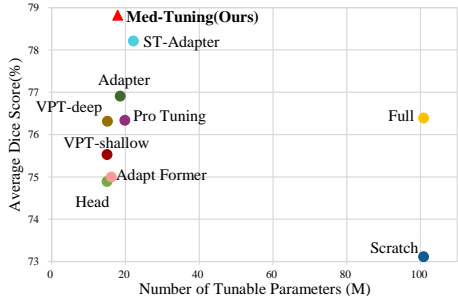
simultaneously considered multi-scale spatial features and 3D volumetric information modeling which are crucial for segmentation. Different from previous works, our Med-Tuning shifts the concentration from classification to dense prediction task (i.e. medical volumetric segmentation) and makes tailored structural design for exploitation of spatial and temporal correlations, realizing the promising PETL with greatly boosted model performance. *It is noteworthy that, despite the recent surge in popularity of the Segment Anything Model (SAM), a few very recent studies such as (Wu et al., 2023; Chai et al., 2023) have focused on leveraging parameter-efficient transfer learning strategies to explore the potential of SAM for medical image analysis. However, the focus of these works primarily rests on SAM, rather than concentrating on the universality and adaptability across various backbones as a plug-and-play component, which is the focal point of our work.*

3 METHODOLOGY

3.1 PRELIMINARIES

Vanilla Adapter. Adapter (Houlsby et al., 2019) is composed of lightweight MLP modules with residual connections and inserted between the feed-forward layer and layer normalization in each Transformer layer. During training, only Adapters are tuned while all the other layers stay frozen. In this way, adapter-based fine-tuning requires much fewer learnable parameters and less training cost than full fine-tuning. Each vanilla adapter utilizes a down-projection linear layer to project the original d-dimensional features into a smaller m-dimension, which is followed by a non-linear activation function and an up-projection linear layer to project features back to d-dimensions. By setting $m \ll d$, the vanilla adapter limits the number of introduced module parameters. Specifically, for the input embedding feature representation $X \in \mathbb{R}^{N \times d}$ from the i-th layer in Transformer, the vanilla adapter can be represented as:

$$\text{Adapter}(\mathbf{X}) = \mathbf{X} + \sigma(\mathbf{X}W_{down})W_{up}, \tag{1}$$

where $W_{down} \in \mathbb{R}^{d \times m}$ and $W_{up} \in \mathbb{R}^{d \times m}$ indicate the down-projection layer and up-projection layer, $\sigma(\cdot)$ is the activation function.

Fourier Transform. Discrete Fourier Transform (DFT) and Inverse Discrete Fourier Transform (IDFT) serve as indispensable techniques for traditional signal analysis, which plays a vital role in our Med-Adapter. Given a sequence data $\mathbf{F} \in \mathbb{R}^N$, a single dimensional DFT $f(k)$ and IDFT $F(n)$ are given below:

$$f(k) = \sum_{n=0}^{N-1} F(n)e^{-j2\pi \frac{kn}{N}}, (k = 0, 1, 2, ..., N-1) \tag{2}$$

$$F(n) = \frac{1}{N} \sum_{x=0}^{N-1} f(k)e^{j2\pi \left(\frac{kn}{N}\right)}, (n = 0, 1, 2, ..., N-1). \tag{3}$$

Furthermore, a 3-dimensional DFT can be computed by the composition of a sequence of one-dimensional DFTs along each dimension (Cooley & Tukey, 1965). Given a 3D data (one image or feature cube) $\mathbf{F} \in \mathbb{R}^{D \times H \times W}$, its 3D-DFT $f(x, y, z)$ and 3D-IDFT $F(d, h, w)$ can be defined as:

$$f(x, y, z) = \sum_{w=0}^{W-1} \sum_{h=0}^{H-1} \sum_{d=0}^{D-1} F(d, h, w)e^{-j2\pi \left(\frac{xd}{D} + \frac{yh}{H} + \frac{zw}{W}\right)}, \tag{4}$$

$$F(d, h, w) = \frac{1}{DHW} \sum_{z=0}^{W-1} \sum_{y=0}^{H-1} \sum_{x=0}^{D-1} f(x, y, z)e^{j2\pi \left(\frac{xd}{D} + \frac{yh}{H} + \frac{zw}{W}\right)}. \tag{5}$$

Note that the accelerated version of DFT and IDFT are employed in our implementation and referred to as FFT and IFFT. When processing 3D images or features with 3D-FFT, the acquired representation is composed of the entire 3D spatial frequency component. Since the FFT operation essentially discretizes spatial domain content into individual frequency components in the frequency domain, each frequency component in the resulting Fourier spectrum has the intrinsic global vision (Rao et al., 2021), which is fully exploited in the global dependency modeling design of our Med-Adapter.

3.2 MEDICAL ADAPTER

In this work, we propose a task-oriented and simple yet effective module for medical data, namely **Med-Adapter**. The PETL framework for various visual Transformer-based models integrated with our Med-Adapters is referred to as **Med-Tuning**.
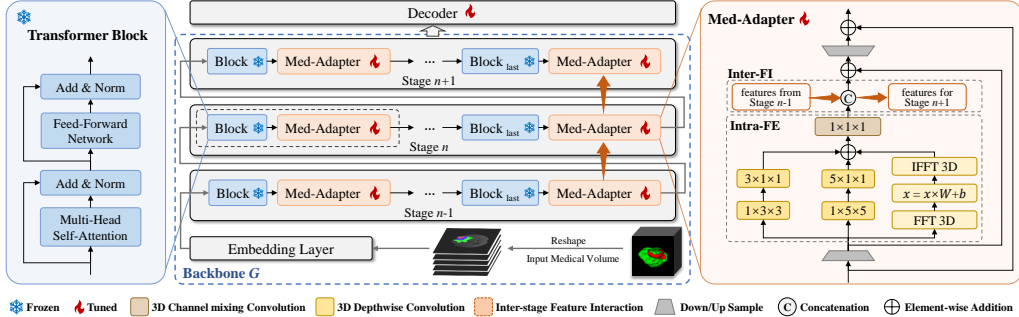
Figure 3: Med-Tuning is our proposed PETL framework, which consists of 2D Transformer baselines for medical volumetric segmentation with our proposed Med-Adapter modules gradually inserted in each stage. Note that we need to reshape and shuffle the 3D input medical volumes from $[B, D, H, W]$ to $[BD, H, W]$ before feeding them to this pipeline, where B = Batch. During training, only Med-Adapters and Decoder are **tuned** while all the other layers are **frozen**. The concatenation operation is calculated along the channel dimension.

The inspiration of our Med-Adapter is to empower a 2D Transformer model pre-trained on natural images to gain the capability of spatial and temporal feature modeling among medical volumes in a parameter-efficient manner. Here we consider three criteria when designing the adapter: (1) *Medical volumetric segmentation task oriented*: The focus of our study is efficiently and effectively narrowing the two-fold gaps mentioned in Sec.1. (2) *Light-weight*: Structure with a low amount of parameters is a typical standard for PETL methods. (3) *Plug-and-play*: An easy-to-implement module is friendly to the practical deployment.

Based on the above inspirations, our Med-Adapter is shown in the right part of Fig. 3. While retaining the overall bottleneck structure of the vanilla adapter (Eq. 1) with a reduction ratio α, a few tailored designs for medical volumetric segmentation are introduced into the internal structure. Formally, given the embedded feature representation $X \in \mathbb{R}^{BD \times C \times HW}$ in Transformer (B, C, D, H, W denote the number of batch size, channel, slice, height, and width respectively), a down-projection layer is first adopted to reduce the embedding dims of tokens, followed by an activation function and a reshape operation to obtain $X' \in \mathbb{R}^{B \times \frac{C}{\alpha} \times D \times H \times W}$, which is expressed as:

$$\mathbf{X}' = \text{Reshape}(\sigma(\mathbf{X}W_{down})), \tag{6}$$

where W_{down} denotes the down-projection layer, $\sigma(\cdot)$ is the activation function.

Intra-stage Feature Enhancement (Intra-FE). Since accurately accomplishing the segmentation task relies on both fine-grained feature representations as well as coarse-grained global semantics, 3D convolutions with diverse kernel sizes are employed to capture the multi-scale representations. Simultaneously, the normal 3D convolution operations are replaced with 3D depth-wise convolutions (Ye et al., 2018) to model the 3D volumetric information in a parameter-efficient manner. Moreover, to pursue an extremely light-weight structure, we take advantage of the combination of $1 \times K \times K$ and $K \times 1 \times 1$ 3D convolutions as an approximation of conventional $K \times K \times K$ 3D convolution (where K denotes the kernel size). As for the frequency branch to realize global dependency modeling, the conventional large-size convolutional kernel and attention mechanism with large memory and computation costs are substituted by parameter-efficient 3D FFT and matrix calculation. In this manner, channel-separable multi-scale features are fully captured by the three parallel branches, followed by a $1 \times 1 \times 1$ convolution to realize efficient channel mixing and obtain the expected layer-wise enhanced feature representation $\mathbf{H}$ with rich multi-scale information. Formally, Intra-FE can be formulated as:

$$\mathbf{F} = IFFT(W_F \odot FFT(\mathbf{X}') + b_F), \tag{7}$$

$$\mathbf{H} = Conv_{1 \times 1 \times 1}(DWConv_3(\mathbf{X}') + DWConv_5(\mathbf{X}') + \mathbf{F}), \tag{8}$$

where FFT and $IFFT$ denote the Fast Fourier Transform and Inverse Fast Fourier Transform, $\odot$ is the Hadamard product, W_F, and b_F are the introduced learnable parameters. $DWConv_K$ denotes two cascaded 3D depth-wise convolutions with the kernel size of $1 \times K \times K$ and $K \times 1 \times 1$.

In this way, our Med-Adapter can effectively and efficiently perform modeling of correlations among slices and capture abundant spatial multi-scale features for the downstream dense prediction task, i.e. medical volumetric segmentation.

Inter-stage Feature Interaction (Inter-FI). Besides, we further consider the feature interaction between different stages. As for the specific Med-Adapters located at the end of each stage, to fully exploit the representations collected by our Med-Adapter at each stage, the intra-stage enhanced feature representation $\mathbf{H}$ will be directly fused with the previous $\mathbf{H}_{LastStage}$ from the corresponding Med-Adapter at the former stage. In this way, feature representations extracted by multiple Med-Adapters in shallow layers are gradually fed to adjacent higher layers, realizing inter-stage feature interaction by explicit enhancement for boosted model performance. Inter-FI is expressed as Eq. 9.

$$\mathbf{H} = \begin{cases} Cat(\mathcal{A}(\mathbf{H}, \mathbf{H}_{LastStage})), & \text{if } last \\ \mathbf{H}, & \text{if not } last \end{cases} \tag{9}$$

where $\mathcal{A}$ denotes using convolutions to realize the alignment between $\mathbf{H}$ and $\mathbf{H}_{LastStage}$ in terms of spatial resolution and channel dimension, Cat refers the concatenation along the channel dimension. $last$ is a bool parameter and $last = True$ when the current Med-Adapter is the last one at stage n.

In summary, our Med-Adapter can be formulated as Eq. 10. $\mathbf{H}$ and $\mathbf{X}'$ are combined by element addition, and then the aggregated feature is symmetrically reshaped back to the same shape as $\mathbf{X}$, followed by the up-projection layer W_{up} and the activation function.

$$\text{Med-Adapter}(\mathbf{X}) = \mathbf{X} + \sigma(\text{Reshape}(\mathbf{H} + \mathbf{X}')W_{up}). \tag{10}$$

3.3 Adapting 2D Transformers to Medical Volumes

The overall architecture of our method, namely Med-Tuning, consists of a commonly utilized decoder and a 2D Transformer backbone G pre-trained on large-scale natural images. As shown in Fig. 3, G has N stages and the n-th stage ($n = 1, 2, ..., N$) has a specific number of Transformer blocks, our proposed Med-Adapters are integrated right after each Transformer block, which makes it friendly for practical deployment as a **plug-and-play** component. Given a batch of 3D medical volume as input $X_B \in \mathbb{R}^{B \times C \times D \times H \times W}$, we first need to reshape them to $X'_B \in \mathbb{R}^{(B \times D) \times C \times H \times W}$ and then send them into the 2D pre-trained backbone. Similarly, the output of the decoder should be reshaped back to the same size as X_B to ensure the alignment of prediction and ground truth. During training, the backbone network is frozen, while only the parameters of our Med-Adapter and the traditional decoder are updated on corresponding datasets.

Through layer-wise insertion and feature interaction between different stages, Med-Adapters obtain and fuse the feature representations with diverse levels. Besides, since our Med-Adapter is not restricted to any specific model structure, Transformer-based architectures can incorporate our framework to greatly reduce the training costs and simultaneously boost model performance.

4 Experiments and Results

4.1 Experimental Setup

Datasets and Evaluation Metrics. Our proposed Med-Tuning is comprehensively evaluated on three benchmark datasets with CT and MRI as two different modalities (i.e. BraTS 2019 (Menze et al., 2014; Bakas et al., 2017; 2018), BraTS 2020 (Menze et al., 2014; Bakas et al., 2017; 2018) and KiTS 2019 (Heller et al., 2019)) for medical image segmentation. Due to space limits, more detailed elaborations are presented in the Appendix A.1.

Implementation Details. Our Med-Tuning framework is implemented based on PyTorch (Paszke et al., 2019) and trained with NVIDIA GeForce RTX 3090 GPUs. As the most representative Transformer-based baselines for medical image segmentation take Swin Transformer (Liu et al., 2021) and Vision Transformer (Dosovitskiy et al., 2020) as visual backbone, Swin-UNet (Cao et al., 2021) and ViT (Dosovitskiy et al., 2020) with UPerNet (Xiao et al., 2018) pre-trained on ImageNet-1k and ImageNet-21k are selected as two strong baselines for a fair comparison. All methods share the same settings with the Adam optimizer during fine-tuning, while the "scratch" version is trained with random initialization (i.e. without any pre-training). More implementation details can be found in the Appendix A.1. The code will be made publicly available.

4.2 Results and Analysis

BraTS 2019. We conduct experiments on the BraTS 2019 validation set and compare our method with previous state-of-the-art (SOTA) approaches for PETL. With the combination of ViT (Doso-

Table 1: Performance comparison on BraTS 2019 and BraTS 2020 with ViT-B/16 backbone pretrained on ImageNet-21k dataset. Red and Blue text denote the performance improvement and the percentage of tuned parameters compared to Full (i.e. full fine-tuning, shown with grey background).

ViT + UPerNet	Tuned Params (M)	Inserted Params (M)	BraTS2019 Dice (%) ↑ ET	WT	TC	Hausdorff (mm) ↓ ET	WT	TC	BraTS2020 Dice (%) ↑ ET	WT	TC	Hausdorff (mm) ↓ ET	WT	TC
Scratch	100.849	-	64.96	83.03	71.34	7.635	10.602	10.942	65.80	83.72	72.01	32.475	10.060	21.467
Full	100.849	-	68.49	85.56	75.12	6.672	7.878	10.525	69.12	85.90	75.29	34.428	7.315	17.093
Head	15.007	-	65.71	84.19	74.77	6.128	7.505	7.864	66.03	84.50	74.47	37.805	7.474	14.150
VPT-Shallow (Jia et al., 2022)	15.015	0.008	66.02	84.72	75.84	6.114	7.506	8.471	66.52	84.82	75.46	37.765	7.465	13.531
VPT-Deep (Jia et al., 2022)	15.100	0.092	67.01	85.14	76.80	6.064	7.717	7.648	67.69	85.28	76.59	31.772	7.737	**10.621**
Adapter (Houlsby et al., 2019)	18.567	3.560	68.30	85.37	77.05	**5.501**	7.636	7.986	68.58	85.77	77.00	32.626	8.172	16.183
AdaptFormer (Chen et al., 2022)	16.197	1.190	65.88	84.34	74.77	6.652	8.204	8.430	65.52	84.14	74.28	41.026	8.393	14.778
Pro-tuning (Nie et al., 2022)	19.812	4.805	67.18	85.32	76.51	5.805	7.073	7.564	67.28	85.57	76.58	40.434	7.000	12.865
ST-Adapter (Pan et al., 2022)	22.118	7.110	69.18	86.27	79.18	6.077	6.939	**6.778**	68.60	86.55	**79.52**	34.060	6.790	12.770
Ours	17.853 (17.70%)	2.846 (2.82%)	**70.53** (+2.04)	**86.58** (+1.02)	**79.35** (+4.23)	5.862 (-0.810)	**6.224** (-1.654)	6.947 (-3.578)	**70.69** (+1.57)	**86.69** (+0.79)	79.36 (+4.07)	**28.643** (-5.785)	**6.198** (-1.117)	15.045 (-2.048)

vitskiy et al., 2020) and UPerNet (Xiao et al., 2018) as the baseline, the comparisons with SOTA methods are presented in Table 1 (left), which shows that our method surpasses most of the previous methods. Additionally, our Med-Tuning also achieves high parameter efficiency, with only 17.70% tuned parameters of the full fine-tuning and inserted parameters that are only 2.82% of fine-tuning all parameters. Compared with other PETL methods, Med-Tuning attains a much better trade-off between performance and efficiency with our Med-Adapters as the easily inserted plug-and-play component, achieving comparable or even better results with smaller parameter costs. Qualitative results are shown in Fig. 4, with comparison to full fine-tuning, ST-Adapter (Pan et al., 2022) and VPT (Jia et al., 2022). As the labels for the validation set are not available, five-fold cross-validation is conducted on the training set for visualization. It can be seen that our method recognizes brain tumors about their enhancing and non-enhancing regions more accurately and reduces missed or false identification of the peritumoral edema in general.

BraTS 2020. We also evaluate our Med-Tuning on BraTS 2020 validation set. As shown in Table 1 (right), with the combination of ViT (Dosovitskiy et al., 2020) and UPerNet (Xiao et al., 2018) as a baseline, our method achieves considerable performance gain on all the metrics compared to full fine-tuning. Compared with previous PETL methods that originated on natural images, Med-Tuning shows better segmentation results while maintaining high parameter efficiency.

Figure 4: The visual comparison of segmentation results on BraTS 2019 dataset. The blue, red and green regions denote the enhancing tumors, non-enhancing tumors, and peritumoral edema. Full and GT denote full fine-tuning and ground truth.

KiTS 2019. To evaluate the generalization ability of our method, we conduct experiments of kidney tumor segmentation on CT scans from the KiTS 2019 dataset (Heller et al., 2019), as shown in Table 2. We can see that the proposed method boosts the performance of full fine-tuning considerably and achieves much higher Dice scores than previous SOTA methods, with much fewer tuned model parameters. In comparison with recently proposed PETL methods (e.g. VPT (Jia et al., 2022), Pro-tuning (Nie et al., 2022) and ST-Adapter (Pan et al., 2022)), our Med-Tuning achieves better performance-efficiency trade-off on two baselines. Specifically, Med-Tuning improves model performance by a large margin (i.e. ↑ **1.01%** Kidney Dice, ↑ **8.02%** Tumor Dice, ↑ **4.52%** Composite Dice on Swin-UNet (Cao et al., 2021) and ↑ **4.20%** Kidney Dice, ↑ **17.13%** Tumor Dice, ↑ **10.67%** Composite Dice on ViT (Dosovitskiy et al., 2020)) with only **27.58%** and **17.70%** of tuned parameters respectively in comparison with full fine-tuning. In addition, the corresponding qualitative comparison presented in the Appendix A.3 shows that our method segments the organs and different kinds of tumors more accurately and generates much better fine-grained segmentation masks of corresponding tumors.
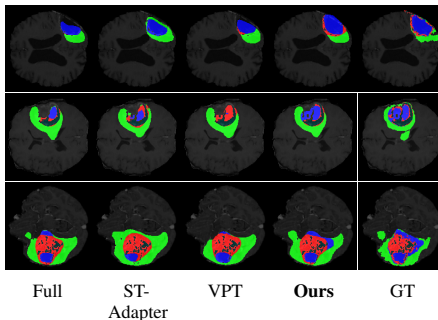
Table 2: Performance comparison on KiTS 2019 with Swin-T backbone pre-trained on ImageNet-1k and ViT-B/16 backbone pre-trained on ImageNet-21k respectively.

Swin-UNet	Tuned Params(M)	Inserted Params(M)	Dice (%) ↑ Kidney	Tumor	Composite	ViT + UPerNet	Tuned Params(M)	Inserted Params(M)	Dice (%) ↑ Kidney	Tumor	Composite
Scratch	27.154	-	94.33	61.10	77.71	Scratch	100.849	-	88.01	46.53	67.27
Full	27.154	-	94.68	62.13	78.40	Full	100.849	-	87.32	47.34	67.33
Head	6.752	-	91.95	53.93	72.94	Head	15.007	-	87.35	42.85	65.10
VPT-Shallow (Jia et al., 2022)	6.753	0.001	91.72	54.86	73.29	VPT-Shallow	15.015	0.008	86.91	41.67	64.29
VPT-Deep (Jia et al., 2022)	6.780	0.029	91.53	53.41	72.47	VPT-Deep	15.100	0.092	88.01	46.45	67.23
Adapter (Houlsby et al., 2019)	7.541	0.790	93.02	57.15	75.08	Adapter	18.567	3.560	89.75	49.03	69.39
AdaptFormer (Chen et al., 2022)	7.124	0.372	93.74	59.79	76.77	AdaptFormer	16.197	1.190	87.62	44.46	66.04
Pro-tuning (Nie et al., 2022)	8.359	1.607	90.34	51.19	70.77	Pro-tuning	19.812	4.805	89.44	48.32	68.88
ST-Adapter (Pan et al., 2022)	8.328	1.577	92.97	57.33	75.15	ST-Adapter	22.118	7.110	90.33	61.29	75.81
Ours	7.489 (27.58%)	0.738 (2.72%)	**95.69** (+1.01)	**70.14** (+8.02)	**82.92** (+4.52)	**Ours**	17.853 (17.70%)	2.846 (2.82%)	**91.52** (+4.20)	**64.47** (+17.13)	**78.00** (+10.67)

4.3 ABLATION STUDIES

We conduct extensive ablation experiments to justify the proposed design based on five-fold cross-validation evaluations on the BraTS 2019 training set. Due to space limits, more ablation studies can be found in the Appendix A.2.

Table 3: Ablation study on our Intra-FE. $Conv_K$ denotes two cascaded 3D depth-wise convolutions with a kernel size of $1 \times K \times K$ and $K \times 1 \times 1$ separately, CM indicates the channel mixing operation by a $1 \times 1 \times 1$ convolution.

$Conv_3$	$Conv_5$	FFT	CM	Tuned Params(M)	Inserted Params(M)	Dice (%) ↑ ET	WT	TC
✓				7.550	0.798	75.42	89.77	80.22
✓	✓			7.574	0.823	75.19	89.44	80.89
✓	✓	✓		7.577	0.825	75.30	89.93	**81.93**
✓	✓	✓	✓	7.675	0.924	**77.10**	**90.05**	81.02

Table 4: Ablation study on inter-stage feature interaction. Swin-UNet with Swin-T pre-trained on supervised ImageNet-1k is taken as a baseline.

Method	Tuned Params(M)	Inserted Params(M)	Dice (%) ↑ ET	WT	TC
Intra-only	7.675	0.924	77.10	90.05	81.02
Add	7.896	1.144	75.79	88.99	79.00
Max	7.896	1.144	75.22	89.72	81.41
Concat	7.994	1.243	**77.22**	**90.09**	**81.59**

Table 5: Ablation study on different encoder pre-trained weights.

Pre-trained Weights	Method	Tuned Params(M)	Inserted Params(M)	Dice (%) ↑ ET	WT	TC
Supervised	Full	100.849	-	66.19	84.72	73.92
	Ours	17.853	2.846	**68.27**	**87.22**	**81.63**
CLIP	Full	100.849	-	64.58	84.69	73.31
	Ours	17.853	2.846	**68.05**	**86.29**	**77.34**
MAE	Full	100.849	-	64.86	84.71	73.95
	Ours	17.853	2.846	**66.32**	**85.50**	**78.05**
MoCo v3	Full	100.849	-	65.06	84.30	73.51
	Ours	17.853	2.846	**67.09**	**85.45**	**77.41**
SAM	Full	100.849	-	65.89	85.32	74.05
	Ours	17.853	2.846	**67.64**	**86.10**	**78.33**

Table 6: Ablation study on reduction ratio α. Swin-UNet with Swin-T pre-trained on supervised ImageNet-1k is chosen as a baseline.

Method	Tuned Params(M)	Inserted Params(M)	Dice (%) ↑ ET	WT	TC
$\alpha=2$	10.064	3.313	76.89	90.14	81.92
$\alpha=4$	7.994	1.243	**77.22**	90.09	81.59
$\alpha=6$	7.489	0.738	77.06	**90.28**	**82.71**
$\alpha=8$	7.271	0.520	76.94	89.62	80.74

Multi-scale Branch Design. We first probe into the rationale of the proposed intra-stage feature enhancement in our Med-Adapter. For the default setting, the reduction ratio α is set to 4 without inter-stage feature interaction. Swin-UNet with Swin-T pre-trained on supervised ImageNet-1k is selected as the baseline. As presented in Table 3, the introduction of either $Conv_5$ branch or FFT branch consistently leads to a considerable performance increase. Specifically, with only 0.002M additional tuned parameters, the FFT branch greatly improves the segmentation accuracy, showing the effectiveness and parameter efficiency of our employed FFT branch. Additionally, channel mixing further boosts the performance by a large margin, especially on ET (↑ 1.80%).

Inter-stage Feature Interaction. After investigating the effect of the intra-stage feature enhancement, we further verify the effectiveness of the inter-stage feature interaction, as shown in Table 4. Compared with the intra-only structure (i.e. without the feature connectivity between adjacent Med-Adapters), the model with inter-stage achieves a considerable performance gain with only 0.319M

extra parameters for feature alignment among adjacent stages, showing the effectiveness of our inter-stage interaction. Unlike concatenation which maintains the feature representations of different stages as much as possible, direct addition or taking the maximum value (at each pixel) of neighboring feature maps with diverse semantic levels would unintentionally degrade the original feature representation, resulting in a sharp decrease in segmentation performance.

Encoder Pre-trained Weights. To explore the potential of our Med-Tuning, we also investigate the effect of diverse encoder pre-trained weights taking ViT-B/16 as the backbone. Since the pre-trained weights of ViT (Dosovitskiy et al., 2020) are relatively easy to acquire, supervised learning-based, multi-modal learning based (i.e. CLIP (Radford et al., 2021)), self-supervised learning based (i.e. MAE (He et al., 2022), MoCo v3 (Chen et al., 2021c)) pre-trained weights are all utilized for a comprehensive comparison. Besides, with the recent advancements in the research of visual foundation models, we also include the Segment Anything Model (i.e. SAM) (Kirillov et al., 2023) for a more comprehensive comparison. As is presented in Table 5, given pre-trained weights acquired by different approaches, our Med-Tuning boosts the performance significantly with much fewer tuned parameters compared with full fine-tuning. Based on the supervised learning-based pre-trained weights such as SAM (Kirillov et al., 2023) that are not pre-trained on medical image domain with great domain gaps, our framework can easily adapt the whole structure for effective medical volumetric segmentation task with all the properly inserted Med-Adapters. *With only 17.70% of the tuned parameters of full fine-tuning, our framework consistently improves the segmentation accuracy by a large margin (i.e. Average Dice scores of 2% to 4%), suggesting the effectiveness, the versatility and the powerful parameter-efficiency of our Med-Tuning framework.*

Reduction Ratio in Bottleneck Design. We analyze the effect of different reduction ratios of the bottleneck structure in our Med-Adapter. Note that the reduction ratio α here is a key factor that influences the tuned parameters introduced by our Med-Adapter. Four diverse settings of α are selected. As shown in Table 6, Med-Tuning achieves a promising trade-off between segmentation accuracy and the tuned parameter costs with $\alpha = 6$. On this basis, higher α would cause inferior model performance because of the deteriorated representation capability with limited tuned parameters, while lower α would lead to a certain degree of information redundancy and a sharp increase of tuned parameters, resulting in both decreased segmentation accuracy and high training costs.

Table 7: Ablation study on data efficiency property with pre-trained ViT-B/16.

Dataset Ratio	Method	Memory Cost (GB)↓	Training Time (h)↓	Dice (%) ↑			HF (mm)↓		
				ET	WT	TC	ET	WT	TC
100%	Full	16.55	1.34	68.04	85.74	76.58	6.94	7.28	7.99
100%	Ours	13.53	1.20	**75.46**	**86.80**	**86.24**	**3.78**	6.94	**4.34**
75%	Ours	13.53	1.05	69.12	86.69	78.06	6.33	**6.01**	6.63
50%	Ours	13.53	0.72	69.19	86.26	77.26	6.28	7.03	7.12
25%	Ours	13.53	0.39	67.43	85.64	74.57	6.32	7.71	8.14
5%	Ours	13.53	0.17	59.61	80.44	64.01	15.07	16.64	16.36

Data Efficiency. At last, we also explore the data efficiency property of our method by examining performance across various training data ratios, particularly in low-data settings. Table 7 shows the quantitative comparison with different numbers of training samples. Our Med-Tuning can already achieve comparable performance to full fine-tuning using only **25%** training data. As the scale of training data increases, our method consistently improves the segmentation accuracy, with reduced training time and memory cost compared with full fine-tuning.

5 CONCLUSION

In this work, we present the study on exploring the potential of PETL for medical volumetric segmentation and propose a new framework named Med-Tuning with high parameter efficiency. Taking advantage of both spatial multi-scale features and 3D volumetric correlations along slices, our framework achieves a trade-off between accuracy and the number of tuned parameters. Extensive experiments show that our method achieves promising performance with greatly shrunk-tuned parameters on three benchmark datasets compared to full fine-tuning and previous PETL SOTA methods.

Our approach provides a new solution of PETL for the practical application of medical volumetric segmentation, suggesting new research in this direction. To some extent, our framework can get rid of the dilemma that the pre-trained weights on large-scale datasets cannot be obtained in the area of medical image analysis. The presented study also encourages the community to consider shifting from constructing large-scale medical image datasets (or pre-training methods) to studying the PETL of pre-trained models on relatively easily acquired natural images.

REFERENCES

Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 1(3):4, 2022.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4: 170117, 2017.

Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021.

Shurong Chai, Rahul Kumar Jain, Shiyu Teng, Jiaqing Liu, Yinhao Li, Tomoko Tateyama, and Yen-wei Chen. Ladder fine-tuning approach for sam integrating complementary network. *arXiv preprint arXiv:2306.12737*, 2023.

Bingzhi Chen, Yishu Liu, Zheng Zhang, Guangming Lu, and David Zhang. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. *arXiv preprint arXiv:2107.05274*, 2021a.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021b.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848, 2017a.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017b.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *arXiv preprint arXiv:2205.13535*, 2022.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9620–9629, 2021c.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pp. 424–432. Springer, 2016.

James W. Cooley and John W. Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of Computation*, 19:297–301, 1965.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. *arXiv preprint arXiv:2209.08575*, 2022.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*, 2021.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*, 2019.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *ICML*, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Xiaohong Huang, Zhifang Deng, Dandan Li, and Xueguang Yuan. Missformer: An effective medical image segmentation transformer. *arXiv preprint arXiv:2109.07162*, 2021.

Yuankai Huo, Jiaqi Liu, Zhoubing Xu, Robert L Harrigan, Albert Assad, Richard G Abramson, and Bennett A Landman. Robust multicontrast mri spleen segmentation for splenomegaly using multi-atlas segmentation. *IEEE Transactions on Biomedical Engineering*, 65(2):336–343, 2017.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*, 2022.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Jiangyun Li, Wenxuan Wang, Chen Chen, Tianxiang Zhang, Sen Zha, Hong Yu, and Jing Wang. Transbtsv2: Wider instead of deeper transformer for medical image segmentation. *arXiv preprint arXiv:2201.12785*, 2022.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34 (10):1993–2024, 2014.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571. IEEE, 2016.

Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *International MICCAI Brainlesion Workshop*, pp. 311–320. Springer, 2018.

Huu-Giao Nguyen, Celine Fouard, and Jocelyne Troccaz. Segmentation, separation and pose estimation of prostate brachytherapy seeds in ct images. *IEEE Transactions on Biomedical Engineering*, 62(8):2012–2024, 2015.

Xing Nie, Bolin Ni, Jianlong Chang, Gaomeng Meng, Chunlei Huo, Zhaoxiang Zhang, Shiming Xiang, Qi Tian, and Chunhong Pan. Pro-tuning: Unified prompt tuning for vision tasks. *ArXiv*, abs/2207.14381, 2022.

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. St-adapter: Parameter-efficient image-to-video transfer learning for action recognition. *arXiv preprint arXiv:2206.13559*, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. *Advances in Neural Information Processing Systems*, 34:980–993, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Lst: Ladder side-tuning for parameter and memory efficient transfer learning. *Advances in Neural Information Processing Systems*, 35:12991–13005, 2022.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20730–20740, 2022.

Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. *arXiv preprint arXiv:2102.10662*, 2021.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021a.

Wenxiao Wang, Lu Yao, Long Chen, Binbin Lin, Deng Cai, Xiaofei He, and Wei Liu. Crossformer: A versatile vision transformer hinging on cross-scale attention. *arXiv preprint arXiv:2108.00154*, 2021b.

Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. Transbts: Multimodal brain tumor segmentation using transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 109–119. Springer, 2021c.

Junde Wu, Rao Fu, Huihui Fang, Yuanpei Liu, Zhaowei Wang, Yanwu Xu, Yueming Jin, and Tal Arbel. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023.

Yixuan Wu, Kuanlun Liao, Jintai Chen, Jinhong Wang, Danny Z Chen, Honghao Gao, and Jian Wu. D-former: A u-shaped dilated transformer for 3d medical image segmentation. *Neural Computing and Applications*, pp. 1–14, 2022.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 418–434, 2018.

Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2945–2954, 2023.

Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. Aim: Adapting image models for efficient video action recognition. *arXiv preprint arXiv:2302.03024*, 2023.

Rongtian Ye, Fangyu Liu, and Liqiang Zhang. 3d depthwise convolution: Reducing model parameters in 3d vision tasks. *ArXiv*, abs/1808.01556, 2018.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

Bruce XB Yu, Jianlong Chang, Lingbo Liu, Qi Tian, and Chang Wen Chen. Towards a unified view on visual parameter-efficient transfer learning. *arXiv preprint arXiv:2210.00788*, 2022.

Ji Zhang, Shihan Wu, Lianli Gao, Hengtao Shen, and Jingkuan Song. Dept: Decoupled prompt tuning. *arXiv preprint arXiv:2309.07439*, 2023a.

Yuhui Zhang, Shihua Huang, Zhengping Zhou, Matthew P. Lungren, and Serena Yeung. Adapting pre-trained vision transformers from 2d to 3d through weight inflation improves medical image segmentation. *ArXiv*, abs/2302.04303, 2023b. URL https://api.semanticscholar.org/CorpusID:256697277.

Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pp. 3–11. Springer, 2018.

A APPENDIX

The appendix is organized into the following sections:

1. (Sec. 1) More specific details about the datasets' information, evaluation metrics and implementation details on the utilized three benchmark datasets (i.e. BraTS 2019, BraTS 2020 and KiTS 2019).
2. (Sec. 2) More experimental results and analysis on BraTS 2019 and BraTS 2020 for a comprehensive investigation.
3. (Sec. 3) More visual comparison of brain tumor segmentation and kidney tumor segmentation for qualitative analysis.

Our code will be publicly available.

A.1 DATASETS, EVALUATION METRICS AND IMPLEMENTATION DETAILS

Datasets and Evaluation Metrics. *Brain Tumor Segmentation 2019 (BraTS 2019)*: The BraTS 2019 (Menze et al., 2014; Bakas et al., 2017; 2018) dataset contains 335 patient cases for training and 125 cases for validation. Each sample consists of 3D brain **MRI** scans with four modalities, while each modality has a volume of $240 \times 240 \times 155$ that has already been aligned into the same space. The ground truth contains 4 classes: background (label 0), necrotic and non-enhancing tumor (label 1), peritumoral edema (label 2), and GD-enhancing tumor (label 4).

Brain Tumor Segmentation 2020 (BraTS 2020): The BraTS 2020 (Menze et al., 2014; Bakas et al., 2017; 2018) dataset's information is identical to BraTS 2019 except for the number of total samples in the dataset. It contains 369 cases for training and 125 cases for validation respectively. On these above two datasets, the segmentation accuracy is measured by Dice score and the Hausdorff distance (95%) metrics for enhancing tumor region (ET, label 4), regions of the tumor core (TC, labels 1 and 4), and the whole tumor region (WT, labels 1,2 and 4).

Kidney Tumor Segmentation 2019 (KiTS 2019): The KiTS 2019 (Heller et al., 2019) dataset is composed of multi-phase 3D **CT**s, including 300 patient cases with high-quality annotated voxel-wise labels. It contains 210 patient cases as the training set and the remaining 90 patients as the testing set. Each CT image/label has a spatial resolution of 512×512 with roughly 50 annotated slices depicting the kidneys and tumors for each case. The ground truth contains 3 classes: background (label 0), kidney (label 1), and kidney tumor (label 2). The same evaluation metrics as KiTS 2019 challenge are utilized: kidney dice considers both kidneys and tumors as foreground, tumor dice considers everything except the tumors as background, and composite dice is the average of kidney dice and tumor dice.

Implementation Details. As shown in Table 8, the specific implementation details on BraTS 2019, BraTS 2020, and KiTS 2019 datasets for two baselines are comprehensively illustrated. On all three benchmark datasets, models are fine-tuned with a batch size of 16 and the Adam optimizer. During training, the following data augmentation techniques are applied to BraTS 2019 and BraTS 2020 datasets: (1) random cropping from $240 \times 240 \times 155$ to $128 \times 128 \times 128$ voxels; (2) random mirror flipping across the axial, coronal and sagittal planes by a probability of 0.5; (3) random intensity shift between [-0.1, 0.1] and scale between [0.9, 1.1]. $L2$ Norm is also applied for regularization with a weight decay rate of 10^{-5}. As for the KiTS 2019 dataset, the employed data augmentations follow as the prior work (Isensee et al., 2021).

Table 8: Implementation details on BraTS 2019, BraTS 2020, and KiTS 2019 datasets for two baselines (i.e. Swin-UNet, ViT) with supervised ImageNet weights.

Dataset	Baseline	Backbone	Pre-train	Learning rate	Training epochs	Warm-up epochs
BraTS 2019 & BraTS 2020	Swin-UNet	Swin-T	ImageNet-1k	0.002	250	60
	ViT+UPerNet	ViT-B/16	ImageNet-21k	0.002	250	25
KiTS 2019	Swin-UNet	Swin-T	ImageNet-1k	0.002	500	20
	ViT+UPerNet	ViT-B/16	ImageNet-21k	0.004	500	20

A.2 MORE RESULTS AND ANALYSIS

In this section, to further explore the potential of our parameter-efficient framework and justify the rationale of its design choices, more comprehensive experiments are conducted. (1) Taking Swin-UNet (Cao et al., 2021) as the second baseline, we evaluate and compare our proposed Med-Tuning framework with previous SOTA approaches for PETL on BraTS 2019 and BraTS 2020 validation sets. (2) Besides, as the supplemented ablation study, we further probe into the effect of various designs for the global dependency modeling of our Med-Adapter Block. (3) Furthermore, we also investigate the effectiveness of different designs for decoder in the whole architecture. (4) Finally, we present more experimental results on a more general and large-scale medical volumetric segmentation benchmark Medical Segmentation Decathlon (MSD) dataset with Swin UNETR (Tang et al., 2022) as a 3D baseline model. Noticeably, the ablation experiment is carried out using five-fold cross-validation evaluations on the BraTS 2019 training set.

Table 9: Performance comparison on BraTS 2019 and BraTS 2020 with Swin-T backbone pre-trained on ImageNet-1k dataset. Red and Blue text denote the performance improvement and the percentage of tuned parameters compared to Full (i.e. full fine-tuning) which has a grey background.

Swin-UNet	Tuned Params (M)	Inserted Params (M)	BraTS2019						BraTS2020					
			Dice (%) ↑			Hausdorff (mm) ↓			Dice (%) ↑			Hausdorff (mm) ↓		
			ET	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC
Scratch	27.154	-	78.38	88.59	76.46	6.055	10.651	9.176	78.72	89.12	77.07	7.621	6.977	19.081
Full	27.154	-	78.26	89.56	79.16	4.327	6.149	6.704	79.09	89.87	79.15	9.671	6.029	15.311
Head	6.752	-	78.07	88.68	77.26	5.021	6.697	7.091	78.77	88.66	76.90	**4.893**	8.491	16.064
VPT-Shallow (Jia et al., 2022)	6.753	0.001	77.16	88.30	76.77	5.421	6.154	7.345	77.43	88.23	76.13	7.532	6.070	16.074
VPT-Deep (Jia et al., 2022)	6.780	0.029	77.02	88.65	76.91	5.297	7.088	7.940	78.63	88.80	77.17	8.265	6.227	13.249
Adapter (Houlsby et al., 2019)	7.541	0.790	77.98	89.22	78.02	5.298	6.622	8.490	78.51	89.16	77.71	7.045	6.245	19.090
AdaptFormer (Chen et al., 2022)	7.124	0.372	77.69	88.61	76.83	4.909	6.290	7.885	78.22	88.92	76.40	10.354	6.480	16.900
Pro-tuning (Nie et al., 2022)	8.359	1.607	**78.58**	89.33	78.79	5.273	6.410	8.236	78.77	89.46	78.20	7.306	6.505	**10.542**
ST-Adapter (Pan et al., 2022)	8.328	1.577	78.40	89.54	77.44	4.751	6.013	7.405	78.96	89.54	77.85	7.673	**5.480**	15.525
Ours	7.489 (27.58%)	0.738 (2.72%)	78.51 (+0.25)	**89.68** (+0.12)	**80.44** (+1.28)	**4.003** (-0.294)	**5.517** (-0.632)	**5.756** (-0.948)	79.25 (+0.16)	**90.06** (+0.19)	**80.79** (+1.64)	12.395 (+2.724)	4.410 (-1.619)	11.591 (-3.720)

A.2.1 PERFORMANCE COMPARISON ON BRATS 2019 AND BRATS 2020 WITH SWIN-UNET

To evaluate the generalization capability of our Med-Tuning, we also select Swin-UNet as another baseline to conduct experiments of brain tumor segmentation on MRI scans utilizing BraTS 2019 (Menze et al., 2014; Bakas et al., 2017; 2018) and BraTS 2020 (Menze et al., 2014; Bakas et al., 2017; 2018) datasets. The corresponding performance comparison on BraTS 2019 and BraTS 2020 validation sets are presented in Table 9. It can be clearly seen that, by jointly exploiting the spatial multi-scale feature representation and correlations along slices, our method once again considerably boosts the model performance of the full fine-tuning paradigm and achieves higher segmentation accuracy than previous SOTA methods with much less tuned parameter costs. To be clear, since Swin-UNet (Cao et al., 2021) with the crucial multi-scale feature representations is essentially a more suitable baseline for dense prediction tasks than ViT (Dosovitskiy et al., 2020), the achieved performance gain by our Med-Tuning on Swin-UNet baseline is understandably not as much as that on ViT. Specifically, with only 27.58% of the tuned parameters of full fine-tuning, our Med-Tuning improves the Dice score and the Hausdorff Distance by a large margin, especially on TC (i.e. ↑ 1.28% and 1.64% on BraTS 2019 and BraTS 2020 validation set separately).

A.2.2 ABLATION STUDY ON DESIGN FOR GLOBAL DEPENDENCY MODELING

Table 10: Ablation study on different designs for global dependency modeling. The baseline is Swin-UNet with Swin-T pre-trained on supervised ImageNet-1k. DWConvK denotes depth-wise convolution with a kernel size of K × K.

Method	Tuned Params(M)	Inserted Params(M)	Dice (%) ↑		
			ET	WT	TC
DWConv9	7.837	1.086	76.48	**90.58**	81.10
DWConv11	8.126	1.375	76.82	89.40	80.05
FFT	7.994	1.243	**77.22**	90.09	**81.59**

In order to pursue the effective and parameter-efficient architecture of our proposed Med-Adapter, we also investigate different designs for the global branch in our Med-Adapter block to achieve global dependency modeling. Since convolutional blocks with a large kernel size or self-attention are usually adopted by previous works for global contextual modeling and the baseline Swin-UNet itself consists of plenty of self-attention operations in each local window, we take the depth-wise convolution with a kernel size of 9 and 11 separately to replace our originally employed Fast Fourier Transform (i.e. FFT) branch for a comprehensive comparison. The comparison of the segmentation performance and tuned model parameters is shown in Table 10. It can be obviously noticed that by taking advantage of the parameter-efficient FFT branch for effective long-range context modeling, the architecture with the FFT branch achieves the optimal trade-off between model performance and tuned parameters, reaching the best segmentation accuracy with only 1.243M introduced model

parameters. In contrast, too large kernel size of the employed convolutions (i.e. DWConv11) will result in a burdensome model structure and a large number of tuned parameter costs.

A.2.3 ABLATION STUDY ON DESIGN FOR DECODER

Table 11: Ablation study on decoder design. ViT-B/16 is pre-trained on supervised ImageNet-1k.

Method	Tuned Params(M)	Decoder Params(M)	Dice (%) ↑		
			ET	WT	TC
UPerNet (Default)	19.562	15.095	68.27	87.22	81.63
U-Net	9.269	4.712	67.68	**88.08**	81.72
SETR-MLA	8.347	3.790	68.12	87.91	**81.98**
SETR-Naive	5.004	0.447	**69.11**	86.93	81.71
SETR-PUP	5.200	0.643	68.55	86.51	80.42

Here we explore the effect of different decoder designs in our architecture. Although the backbone is frozen and only the inserted Med-Adapters as well as the decoder are updated during fine-tuning, the essentially tuned model parameters introduced by the segmentation decoder can not be reckoned as negligible. In other words, to pursue an extremely PETL framework, the design of the employed decoder should be sufficiently lightweight with strictly controlled model parameters. Thus, various segmentation decoders with greatly varied model complexity are introduced respectively for a thorough analysis. As shown in Table 11, ViT-B/16 with the SETR-MLA decoder reaches the best trade-off between segmentation accuracy and tuned parameter costs, benefiting from the effective multi-scale feature aggregation. Besides, taking the simplest SETR-Naive that is composed of a convolution and an interpolation operation for upsampling as the decoder leads to the lowest tuned parameters 5.004M while achieving promising segmentation performance with an average Dice score of 79.34%. It can be seen from Table 11 that although the decoder size dominantly decides the overall tuned parameters, it does not show a direct impact on model performance.

A.2.4 RESULTS ON MEDICAL SEGMENTATION DECATHLON (MSD) DATASET

To further evaluate the generalization capability and superiority of our Med-Tuning, we conduct experiments to encompass additional body parts on another popular benchmark Medical Segmentation Decathlon (MSD) dataset. *Simultaneously, as a study on transfer learning in medical volumetric segmentation, since 2D models are naturally outperformed by 3D counterparts with richer temporal information and a few available 3D models pre-trained on medical data exists, the fair comparison with this kind of baselines is essential for our work to achieve a more comprehensive comparison. Therefore, we choose Swin UNETR (Tang et al., 2022) as a supplementary 3D medical pre-trained baseline for comparison to convince the effectiveness of our method over both 3D models and baseline models pre-trained on medical data.* The experimental results of three tasks (i.e. heart, lung, spleen) are listed in Table 12. Same conclusion can be drawn from Table 12 that our method consistently boosts model performance with less memory cost and training time on various body parts.

Table 12: Results on MSD dataset with pre-trained Swin UNETR.

Organ	Method	Memory(GB)↓	Time(h)↓	Dice_AVG(%)↑
Task02 Heart (MRI)	Scratch	19.73	1.05	91.95
	Full	19.73	1.06	93.73
	Ours	**13.44**	**0.86**	**95.84**
Task06 Lung (CT)	Scratch	23.51	8.39	65.82
	Full	23.51	8.39	67.69
	Ours	**20.30**	**8.03**	**78.09**
Task09 Spleen (CT)	Scratch	20.32	3.21	95.76
	Full	20.32	3.21	96.52
	Ours	**19.71**	**2.22**	**97.06**

A.3 MORE VISUAL COMPARISON

To further demonstrate the advantage of our proposed framework, we present more visualizations of brain tumor segmentation and kidney tumor segmentation results on BraTS 2019 dataset and KiTS 2019 dataset respectively for qualitative analysis in Fig. 5 and Fig. 6. The various methods utilized for visual comparison include traditional full fine-tuning paradigm, ST-Adapter (Pan et al., 2022), VPT (Jia et al., 2022) and our Med-Tuning. It can be clearly observed from both Fig. 5 and Fig. 6

that our method can segment the specific organs and different kinds of tumors more accurately and generate much better fine-grained segmentation masks of corresponding tumors.
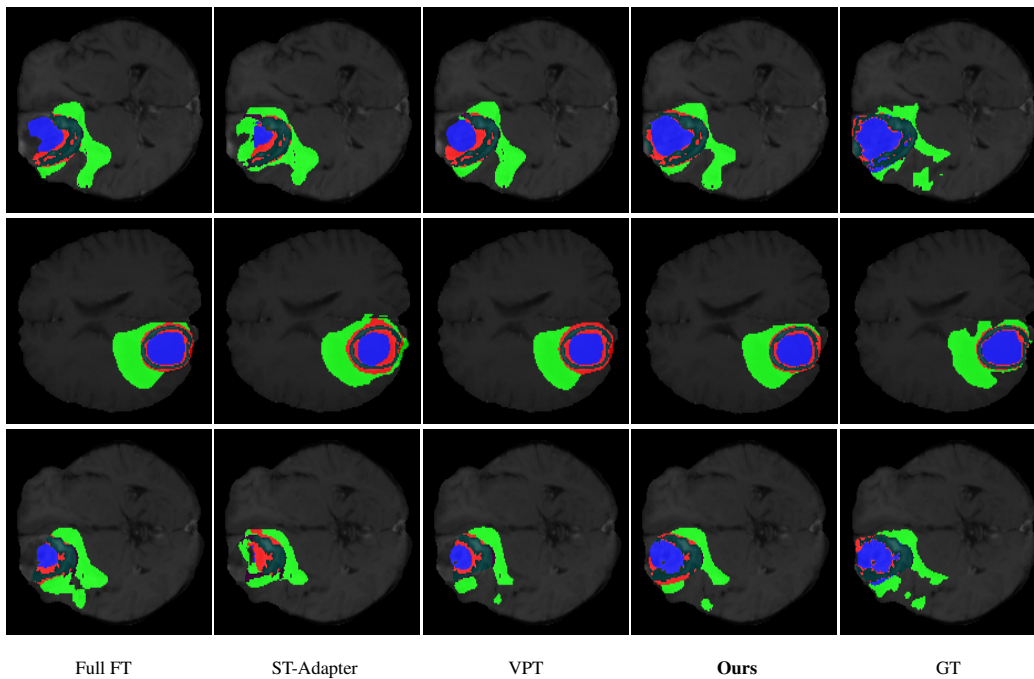

| Full FT | ST-Adapter | VPT | **Ours** | GT |

Figure 5: The visual comparison of MRI brain tumor segmentation results. The blue regions denote the enhancing tumors, the red regions denote the non-enhancing tumors, and the green ones denote the peritumoral edema. Full FT and GT denote full fine-tune and ground truth respectively.
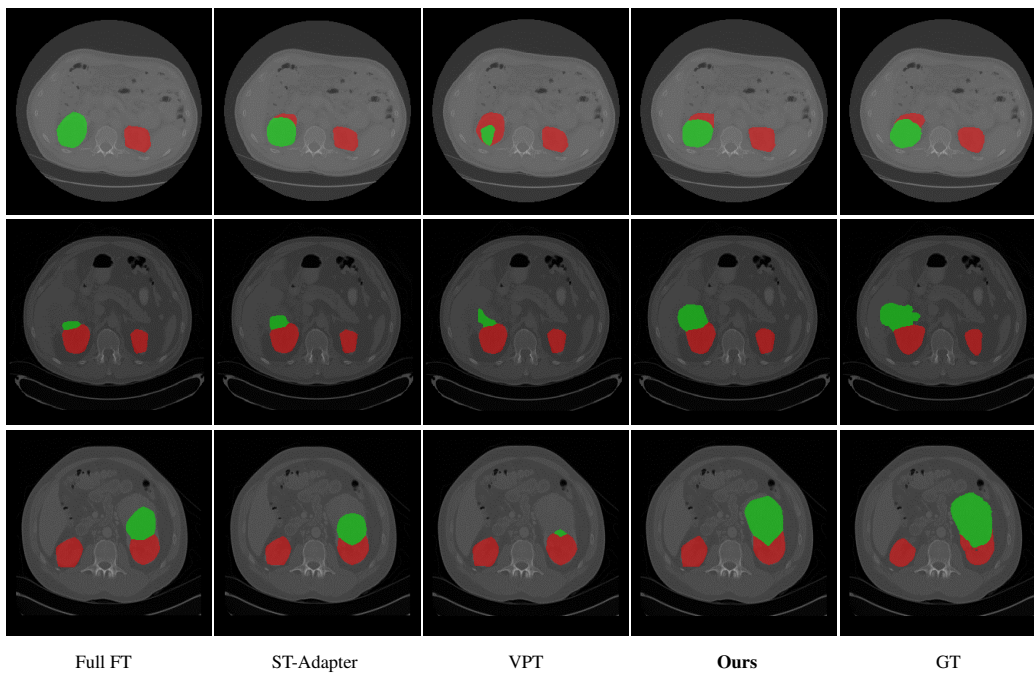
Figure 6: The visual comparison of segmentation results on KiTS 2019 dataset. The red and green regions denote the kidneys and kidney tumors. Full and GT denote full fine-tuning and ground truth.