Med-Tuning

Wang, Wenxuan; Shen, Jiachen; Chen, Chen; Jiao, Jianbo; Zhang, Yan; Song, Shanshan; Li, Jiangyun

DOI:
10.48550/arXiv.2304.10880

License:
Other (please provide link to licence statement

Document Version
Other version

Citation for published version (Harvard):
Wang, W, Shen, J, Chen, C, Jiao, J, Zhang, Y, Song, S & Li, J 2023 'Med-Tuning: Exploring Parameter-Efficient Transfer Learning for Medical Volumetric Segmentation' arXiv. https://doi.org/10.48550/arXiv.2304.10880

Med-Tuning: Exploring Parameter-Efficient Transfer Learning for Medical Volumetric Segmentation

Wenxuan Wang[1,*] Jiachen Shen[1,*] Chen Chen[2] Jianbo Jiao[3]
Yan Zhang[1] Shanshan Song[1] Jiangyun Li[1†]
[1]School of Automation and Electrical Engineering, University of Science and Technology Beijing
[2]Center for Research in Computer Vision, University of Central Florida
[3]School of Computer Science, University of Birmingham
{s20200579,m202110559}@xs.ustb.edu.cn, chen.chen@crcv.ucf.edu
jiaojianbo.i@gmail.com,leejy@ustb.edu.cn

Abstract

Deep learning based medical volumetric segmentation methods either train the model from scratch or follow the standard "pre-training then finetuning" paradigm. Although finetuning a well pre-trained model on downstream tasks can harness its representation power, the standard full finetuning is costly in terms of computation and memory footprint. In this paper, we present the first study on parameter-efficient transfer learning for medical volumetric segmentation and propose a novel framework named Med-Tuning based on intra-stage feature enhancement and inter-stage feature interaction. Given a large-scale pre-trained model on 2D natural images, our method can exploit both the multi-scale spatial feature representations and temporal correlations along image slices, which are crucial for accurate medical volumetric segmentation. Extensive experiments on three benchmark datasets (including CT and MRI) show that our method can achieve better results than previous state-of-the-art parameter-efficient transfer learning methods and full finetuning for the segmentation task, with much less tuned parameter costs. Compared to full finetuning, our method reduces the finetuned model parameters by up to 4×, 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 to make 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) [33] and magnetic resonance imaging (MRI) [24], 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) [40, 31, 51, 32, 12, 53, 35, 22, 25] and Transformers [6, 42, 23, 46, 5, 43, 27]. With the continuous improvement of model performance, the number of model parameters and corresponding training cost have increased greatly, especially the Transformer-based models. Besides, due to the challenges in model training these methods benefit from finetuning the models pre-trained on larger-scale datasets (*e.g.* ImageNet [14]), 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 finetuning parameter efficiency?*

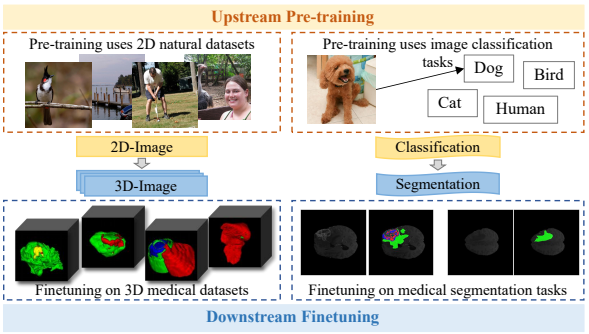


Figure 1. The illustration of the two-fold gaps between source and target domain when exploring pre-trained model on large-scale 2D natural image datasets for medical volumetric segmentation.

In the community of natural image processing, ***"pre-training then finetuning"*** paradigm has become standard

*Equal Contribution. †Corresponding author.

practice to boost the model performance on downstream tasks. Conventional finetuning schemes include *full finetuning* and *head finetuning*, which optimize either the entire network or only the specific head (*e.g*. Linear [18] and Partial [49]). Full finetuning usually achieves higher accuracy but also a higher training cost. Recent studies [26, 34, 10, 36, 50] on parameter-efficient transfer learning (PETL) try to achieve a balance in between.

In this paper, we present, to the best of our knowledge, the first attempt to 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. Therefore, *the objective of this work is to explore how to effectively and efficiently adapt strong pre-trained models on 2D natural images to the medical volumetric segmentation task*.

As shown in Fig. 1, there are two-fold gaps between the pre-training source domain and the downstream target domain that need to be considered 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. In order to narrow these gaps, we propose to build a PETL framework for medical volumetric segmentation based on pre-trained models on natural images with an efficient plug-and-play block to exploit the crucial spatial multi-scale features and temporal correlations.

Specifically, for the first gap brought by 3D medical data itself, there is an essential temporal continuity between adjacent medical image slices that need to be exploited. 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 [29, 41, 7, 8, 9, 44, 28, 16, 45] 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 the first 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.
- Extensive experiments on three benchmark datasets (includes **CT** and **MRI**) validate the effectiveness (*e.g*. Fig. 2) of our Med-Tuning over full finetuning and previous PETL methods for medical volumetric segmentation.

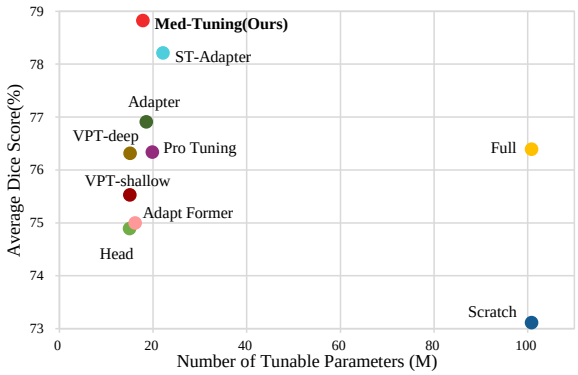


Figure 2. Comparison with previous PETL methods in terms of trade-off between tuned parameters and segmentation accuracy. The backbone ViT-B/16 is pre-trained on ImageNet-21k and fine-tuned on BraTS2019 dataset. Our method achieves much better segmentation performance than full finetuning and previous state-of-the-art PETL methods with much less tuned parameters.

2. Related Work

2.1. Medical Volumetric Segmentation

Unlike natural images, medical images have particular challenges, such as uneven distribution of foreground and background, and sharp changes in shapes and scales of lesions. As proved by previous works [12, 31, 53, 5, 6, 35], extracting multi-scale representations is crucial for desired segmentation performance. For example, Ronneberger *et al*. [40] concatenated the multi-scale features from the CNN encoder and the up-sampled features together, complementing the loss of spatial information caused by down-samplings. Cao *et al*. [4] also used skip connections to gradually fuse the low-level features and the high-level features together in Transformer architecture.

In addition, the information between continuous slices (*i.e*. temporal correlation) of medical volumetric images is of critical importance. Various medical volumetric segmentation methods [12, 31, 53, 35] have effectively taken advantage of this vital continuity by typically utilizing 3D convolutions [12, 31, 53] or introducing self-attention mechanism among the 3D input patches [43].

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 finetuning methods can not 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 [1], which modifies the input pixel space of Transformer layers. VPT [26] prepended a series of learnable prompts to the patch embeddings to facilitate downstream visual tasks. But VPT is sensitive to the number of prompts and token length, which may have limited potential in parameter efficiency for dense prediction tasks. Pro-tuning [34] inserted multiple stage-wise prompt blocks into different stages of the backbone.

The *second* type is Adapter that can be easily inserted into backbones. Specifically, AdaptFormer [10] replaced the original multi-layer perceptron (MLP) block in Transformer with the proposed AdaptMLP. Despite its promising results, AdaptFormer does not take temporal information into account, which may lead to the loss of connections between video clips. To tackle this problem, ST-Adapter [36] injected Adapter-like blocks in each Transformer layer and introduced the 3D depth-wise convolution [48] to capture spatial-temporal features. However, it does not take the modeling of multi-scale representation into consideration, which is critical for the segmentation task.

The *third* category includes other PETL techniques. For example, LoRA [21] inserted learnable low-rank matrices into the self-attention block in Transformer, while V-PETL [50] extended the parameters of prefix tuning [17] from randomly initialized to input associated.

Nevertheless, the above previous researches mainly pay attention to 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 have simultaneously considered multi-scale features and temporal information modeling which are crucial for segmentation. Different from previous works, our Med-Tuning pioneeringly 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.

2.3. Fourier Transform in Deep Learning

Image analysis in the Fourier domain has been widely explored in diverse computer vision tasks. Notably, the Fourier transform utilizes frequency information to naturally build global connectivity by operating domain mapping on original images in a parameter-free manner (*i.e.* without any additional parameters). For instance, GFNet [39] focused on model structure modification and substituted the vanilla self-attention blocks in the original Transformer with FFT operation, realizing efficient global feature modeling on high-resolution images. [52] proposed to upsample in the frequency domain to avoid the inability of exploiting global dependency as common up-sampling (*i.e.* interpolation, transposed convolution, etc.) in the spatial domain. Inspired by the above works, we present the first study on exploiting the intrinsic global properties of FFT for PETL and propose a novel adapter block namely Med-Adapter with a well-designed FFT branch, with the aim of effectively and efficiently modeling the crucial global context for medical volumetric segmentation.

3. Methodology

3.1. Preliminaries

Vanilla Adapter. Adapters [20] are 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 finetuning requires much fewer learnable parameters and less training cost than full finetuning. 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:

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

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

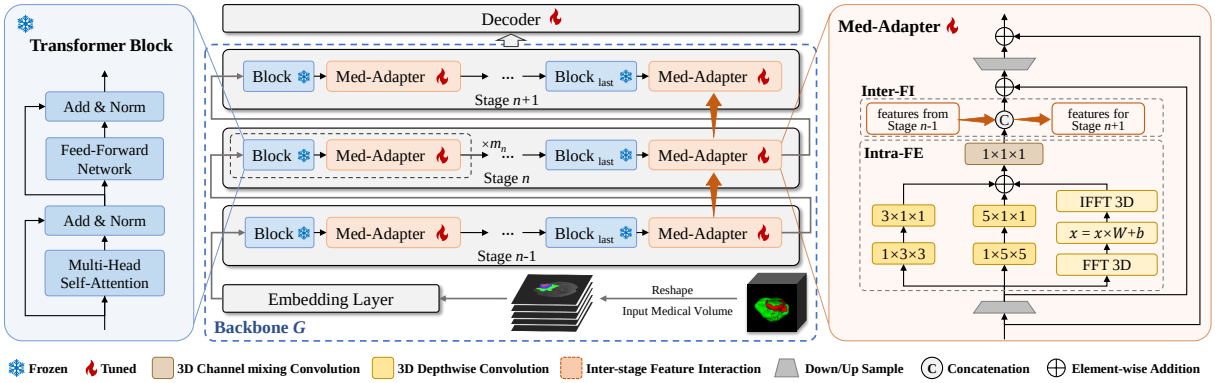


Figure 3. Med-Tuning is a finetuning framework, which consists of 2D Transformer baselines for medical volumetric segmentation with our proposed Med-Adapter modules gradually inserted in each stage. Note that we need to reshape and shuffle the 3D input medical volumes from $[B, D, H, W]$ to $[BD, H, W]$ 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.

Furthermore, a 3-dimension DFT can be computed by the composition of a sequence of one-dimensional DFTs along each dimension [13]. Given a 3D data (one image cube 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:

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

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

Note that the accelerated version of DFT and IDFT are employed in our implementation and referred 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, 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, namely **Med-Adapter**. The PETL framework for vision Transformer integrated with our Med-Adapters is referred to as **Med-Tuning**.

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. Several important criteria of designing should be followed: (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 can be expressed as:

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

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 [48] to model the required temporal information in a parameter-efficient manner. Moreover, for the purpose of pursuing 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:

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

$$\mathbf{H} = Conv_{1\times1\times1}(DWConv_3(\mathbf{X}') + DWConv_5(\mathbf{X}') + \mathbf{F}), \tag{8}$$

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 $K\times1\times1$.

In this way, our Med-Adapter can effectively and efficiently perform modeling of correlations among temporal 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 feature 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. $last$ 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 together by element addition, 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 Medilal 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 $m_n + 1$ 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 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 specific datasets.

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

4. Experiments and Results

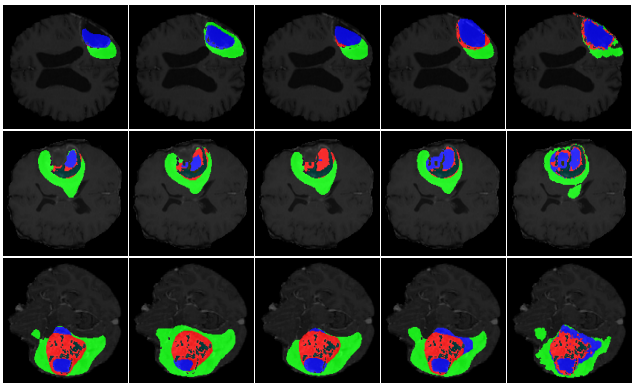
4.1. Experimental Setup

Datasets and Evaluation Metrics.

Brain Tumor Segmentation 2019 (BraTS 2019): The BraTS 2019 [30, 2, 3] 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 [30, 2, 3] 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 [19] 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.



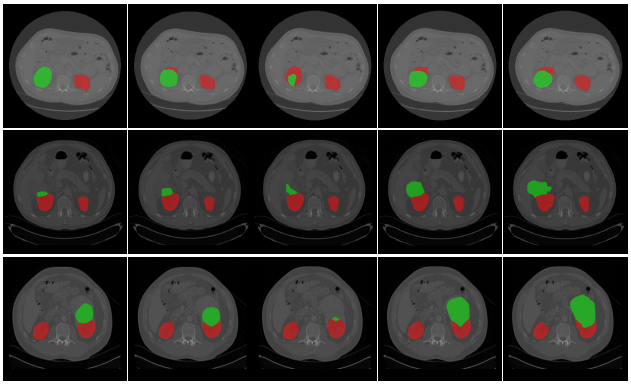
Full ST-Adapter VPT **Ours** GT

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, GT denote full finetuning and ground truth.

Implementation Details. The proposed Med-Tuning framework is implemented based on Pytorch [37] and trained with NVIDIA GeForce RTX 3090 GPUs. As the most representative Transformer-based baselines for medical image segmentation, Swin-UNet [4] and ViT [15] with UPerNet [47] are selected as the Transformer-based baselines with the large-scale pre-trained weights from ImageNet-1k and ImageNet-21k respectively. All methods share the same settings with Adam optimizer during finetuning, while the "scratch" version is trained with random initialization (*i.e.* without any pre-trained weights).

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 [15] and UPerNet [47] as the baseline, the comparisons with state-of-the-art methods are presented in Table 1 (left), which shows that our method surpasses most of the previous methods. Besides of the competitive segmentation performance, our Med-Tuning also achieves high parameter efficiency, with only 17.70% tuned parameters of the full finetuning and inserted parameters that is only 2.82% of finetuning all parameters. Compared with other PETL methods, Med-Tuning attains much better trade-off between performance and efficiency, achieving comparable or even better results with smaller parameter costs. Qualitative results on brain tumor segmentation are shown in Fig. 4, with comparison to full finetuning, ST-Adapter[36] and VPT[26]. 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.



Full ST-Adapter VPT **Ours** GT

Figure 5. The visual comparison of segmentation results on KiTS 2019 dataset. The red and green regions denote the kidneys and kidney tumors. Full, GT denote full finetuning and ground truth.

BraTS 2020. We also evaluate our Med-Tuning on BraTS 2020 validation set. As shown in Table 1 (right), with the combination of ViT [15] and UPerNet [47] as the baseline, our method achieves 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.

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 [19], as shown in Table 2. We can see that the proposed method boosts the performance of full finetuning significantly and achieves much higher Dice scores than previous state-of-the-art methods, with much fewer tuned model parameters. In comparison with recently proposed PETL methods (*e.g.* VPT[26], Pro-tuning[34] and ST-Adapter[36]), 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 [4] and ↑ **4.20%** Kidney Dice, ↑ **17.13%** Tumor Dice, ↑ **10.67%** Composite Dice on ViT [15]) with only **27.58%** and **17.70%** of tuned parameters respectively in comparison with full finetuning. In addition, qualitative comparison in Fig. 5 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.

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 dataset.

Multi-scale Branch Design. We firstly probe into the rationale of the proposed intra-stage feature enhancement in our Med-Adapter. For the default setting, the reduc-

ViT [15]+ UPerNet [47]	Tuned Params (M)	Inserted Params (M)	BraTS2019 Dice (%) ↑ ET	WT	TC	BraTS2019 Hausdorff (mm) ↓ ET	WT	TC	BraTS2020 Dice (%) ↑ ET	WT	TC	BraTS2020 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 [26]	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 [26]	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 [20]	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 [10]	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 [34]	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 [36]	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)

Table 1. Performance comparison on BraTS 2019 and BraTS 2020 with ViT-B/16 pre-trained on ImageNet-21k dataset. Red color denotes performance improvement compared to Full (*i.e.* full finetuning) which has a grey background.

Swin-UNet [4]	Tuned Params(M)	Inserted Params(M)	Dice (%) ↑ Kidney	Tumor	Composite	ViT [15]+ UPerNet [47]	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 [26]	6.753	0.001	91.72	54.86	73.29	VPT-Shallow [26]	15.015	0.008	86.91	41.67	64.29
VPT-Deep [26]	6.780	0.029	91.53	53.41	72.47	VPT-Deep [26]	15.100	0.092	88.01	46.45	67.23
Adapter [20]	7.541	0.790	93.02	57.15	75.08	Adapter [20]	18.567	3.560	89.75	49.03	69.39
AdaptFormer [10]	7.124	0.372	93.74	59.79	76.77	AdaptFormer [10]	16.197	1.190	87.62	44.46	66.0 4
Pro-tuning [34]	8.359	1.607	90.34	51.19	70.77	Pro-tuning [34]	19.812	4.805	89.44	48.32	68.88
ST-Adapter [36]	8.328	1.577	92.97	57.33	75.15	ST-Adapter [36]	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)

Table 2. Performance comparison on KiTS 2019 with Swin-T pre-trained on ImageNet-1k and ViT-B/16 pre-trained on ImageNet-21k respectively. Red text denotes performance improvement compared to Full (*i.e.* full finetuning) which has a grey background.

$Conv_3$	$Conv_5$	FFT	CM	Tuned Params(M)	Inserted Params(M)	Dice (%) ↑ ET	WT	TC	Avg.
✓				7.550	0.798	75.42	89.77	80.22	81.80
✓	✓			7.574	0.823	75.19	89.44	80.89	81.84
✓	✓	✓		7.577	0.825	75.30	89.93	**81.93**	82.39
✓	✓	✓	✓	7.675	0.924	**77.10**	**90.05**	81.02	**82.72**

Table 3. Ablation study on intra-stage feature enhancement. $Conv_K$ 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.

tion ratio α is set to 4 without inter-stage feature interaction. Swin-UNet with Swin-T was pre-trained on supervised ImageNet-1k. 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, FFT branch significantly improves the segmentation accuracy (*i.e.* $\uparrow 1.04\%$ and $\uparrow 0.55\%$ on TC and average Dice respectively), 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 ($\uparrow 1.80\%$) and the average Dice score ($\uparrow 0.33\%$).

Method	Tuned Params(M)	Inserted Params(M)	Dice (%) ↑ ET	WT	TC	Avg.
DWConv9	7.837	1.086	76.48	**90.58**	81.10	82.72
DWConv11	8.126	1.375	76.82	89.40	80.05	82.09
FFT	7.994	1.243	**77.22**	90.09	**81.59**	**82.97**

Table 4. 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.

Design for Global Dependency Modeling. In order to pursue the most 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 operation in each local window, we take the depth-wise convolution with 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 4. It can be obviously noticed that by taking advantage of the parameter-efficient FFT branch for effective long-range context modeling, the architecture with 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 burdensome model structure and large amount of tuned parameter costs.

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 5. Compared with the intra-only structure (*i.e.* without the feature connectivity between adjacent Med-Adapters), the model with inter-stage achieves a consider-

Method	Tuned Params(M)	Inserted Params(M)	Dice (%) ↑ ET	WT	TC	Avg.
Intra-only	7.675	0.924	77.10	90.05	81.02	82.72
Add	7.896	1.144	75.79	88.99	79.00	81.26
Max	7.896	1.144	75.22	89.72	81.41	82.12
Concat	7.994	1.243	**77.22**	**90.09**	**81.59**	**82.97**

Table 5. Ablation study on inter-stage feature interaction. Swin-UNet with Swin-T pre-trained on supervised ImageNet-1k.

Method	Tuned Params(M)	Inserted Params(M)	Dice (%) ↑ ET	WT	TC	Avg.
α=2	10.064	3.313	76.89	90.14	81.92	82.99
α=4	7.994	1.243	**77.22**	90.09	81.59	82.97
α=6	7.489	0.738	77.06	**90.28**	**82.71**	**83.35**
α=8	7.271	0.520	76.94	89.62	80.74	82.44

Table 6. Ablation study on reduction ratio α. Swin-UNet with Swin-T pre-trained on supervised ImageNet-1k.

Pre-trained Weights	Method	Tuned Params(M)	Inserted Params(M)	Dice (%) ↑ ET	WT	TC	Avg.
Supervised	Full	100.849	-	66.19	84.72	73.92	74.94
	Ours	17.853	2.846	**68.27**	**87.22**	**81.63**	**79.04**
CLIP	Full	100.849	-	64.58	84.69	73.31	74.19
	Ours	17.853	2.846	**68.05**	**86.29**	**77.34**	**77.23**
MAE	Full	100.849	-	64.86	84.71	73.95	74.51
	Ours	17.853	2.846	**66.32**	**85.50**	**78.05**	**76.62**
MoCo v3	Full	100.849	-	65.06	84.30	73.51	74.29
	Ours	17.853	2.846	**67.09**	**85.45**	**77.41**	**76.65**

Table 7. Ablation study on different encoder pre-trained weights.

able 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.

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 (*i.e.* 2, 4, 6, 8) of the reduction ratio α are selected. As shown in Table 6, Med-Tuning achieves 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.

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 [15] are relatively easy to acquire, supervised learning-based, multi-modal learning based (*i.e.* CLIP [38]) and self-supervised learning based (*i.e.* MAE [18], MoCo v3 [11]) pre-trained weights are all utilized for a comprehensive comparison.

Method	Tuned Params(M)	Decoder Params(M)	Dice (%) ↑ ET	WT	TC	Avg.
UPerNet (Default)	19.562	15.095	68.27	87.22	81.63	79.04
U-Net	9.269	4.712	67.68	**88.08**	81.72	79.16
SETR-MLA	8.347	3.790	68.12	87.91	**81.98**	**79.34**
SETR-Naive	5.004	0.447	**69.11**	86.93	81.71	79.25
SETR-PUP	5.200	0.643	68.55	86.51	80.42	78.49

Table 8. Ablation study on decoder design. ViT-B/16 is pre-trained on supervised ImageNet-1k.

As is presented in Table 7, given pre-trained weights acquired by different approaches, our Med-Tuning boosts the performance significantly with much fewer tuned parameters compared with full finetuning. With only 17.70% of the tuned parameters of full finetuning, our framework improves the segmentation accuracy by a large margin (*i.e.* Average Dice scores of 2% to 4%), suggesting the effectiveness and the parameter-efficiency of our Med-Tuning.

Decoder Design. 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 finetuning, 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 8, 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 8 that although the decoder size dominantly decides the overall tuned parameters, it does not show a direct impact on model performance.

5. Conclusion

We present to our knowledge the *first study* on exploring the potential of PETL for medical volumetric segmentation task and propose a new framework named Med-Tuning with high parameter efficiency. Taking advantage of both spatial multi-scale feature and temporal correlations, our framework achieves the trade-off between segmentation 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 finetuning and previous PETL SOTA methods.

Our approach provides a novel solution of PETL for the practical application of medical volumetric segmentation,

which inspires 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 and encourage the community to consider shifting the research perspective 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

[1] 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. 3

[2] 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. 5

[3] 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. 5

[4] 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. 2, 6, 7

[5] 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*, 2021. 1, 2

[6] 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*, 2021. 1, 2

[7] 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. 2

[8] 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, 2017. 2

[9] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 2

[10] 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. 2, 3, 7

[11] 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)*, pages 9620–9629, 2021. 8

[12] 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*, pages 424–432. Springer, 2016. 1, 2

[13] James W. Cooley and John W. Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of Computation*, 19:297–301, 1965. 4

[14] 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*, pages 248–255. Ieee, 2009. 1

[15] 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. 6, 7, 8

[16] 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. 2

[17] 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. 3

[18] 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*, pages 16000–16009, 2022. 2, 8

[19] 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. 5, 6

[20] 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. 3, 7

[21] 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. 3

[22] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1055–1059. IEEE, 2020. 1

[23] Xiaohong Huang, Zhifang Deng, Dandan Li, and Xueguang Yuan. Missformer: An effective medical image segmentation transformer. *arXiv preprint arXiv:2109.07162*, 2021. 1

[24] 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. 1

[25] 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. 1

[26] 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. 2, 3, 6, 7

[27] 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. 1

[28] 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. 2

[29] 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*, pages 3431–3440, 2015. 2

[30] 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. 5

[31] 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)*, pages 565–571. IEEE, 2016. 1, 2

[32] Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *International MICCAI Brainlesion Workshop*, pages 311–320. Springer, 2018. 1

[33] 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. 1

[34] 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. 2, 3, 6, 7

[35] 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. 1, 2

[36] 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. 2, 3, 6, 7

[37] 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. 6

[38] 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. 8

[39] 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. 3

[40] 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*, pages 234–241. Springer, 2015. 1, 2

[41] 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*, pages 1–9, 2015. 2

[42] 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. 1

[43] 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*, pages 109–119. Springer, 2021. 1, 2

[44] 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*, 2021. 2

[45] 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*, 2021. 2

[46] 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*, pages 1–14, 2022. 1

[47] 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)*, pages 418–434, 2018. 6, 7

[48] Rongtian Ye, Fangyu Liu, and Liqiang Zhang. 3d depthwise convolution: Reducing model parameters in 3d vision tasks. *ArXiv*, abs/1808.01556, 2018. 3, 4

[49] 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. 2

[50] 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. 2, 3

[51] 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. 1

[52] Man Zhou, Hu Yu, Jie Huang, Feng Zhao, Jinwei Gu, Chen Change Loy, Deyu Meng, and Chongyi Li. Deep fourier up-sampling. *arXiv preprint arXiv:2210.05171*, 2022. 3

[53] 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*, pages 3–11. Springer, 2018. 1, 2