Towards foundation models and few-shot parameter-efficient fine-tuning for volumetric organ segmentation

Julio Silva-Rodríguez, Jose Dolz, and Ismail Ben Ayed

ETS Montreal, Canada

Abstract. With the recent raise of foundation models in computer vision and NLP, the *pretrain-and-adapt* strategy, where a large-scale model is fine-tuned on downstream tasks, is gaining popularity. However, traditional fine-tuning approaches may still require significant resources and yield sub-optimal results when the labeled data of the target task is scarce. This is especially the case in clinical settings. To address this challenge, we formalize few-shot efficient fine-tuning (FSEFT), a novel and realistic setting for medical image segmentation. Furthermore, we introduce a novel parameter-efficient fine-tuning strategy tailored to medical image segmentation, with (a) spatial adapter modules that are more appropriate for dense prediction tasks; and (b) a constrained transductive inference, which leverages task-specific prior knowledge. Our comprehensive experiments on a collection of public CT datasets for organ segmentation reveal the limitations of standard fine-tuning methods in few-shot scenarios, point to the potential of vision adapters and transductive inference, and confirm the suitability of foundation models. The project code is available in `https://github.com/jusiro/fewshot-finetuning`.

Keywords: Efficient fine-tuning · Few-shot adapters · Transduction

1 Introduction

The recent advancements in deep learning have yielded remarkable outcomes in visual recognition tasks. Specifically, under the standard supervised learning paradigm, training on sufficiently large amounts of labeled data could yield excellent performances in medical image segmentation. The success of several recent public challenges, including [1,15,16], attests to this. However, these models are often trained on a specific task and limited numbers of samples, which may lack real-world inter-center variability. As a result, the current literature suggests that general medical image segmentation is hampered by the lack of large, curated datasets for training [8]. This limitation is further exacerbated in volumetric medical image segmentation, where expert knowledge is required for voxel-wise annotations. For example, an experienced clinician would require an average of 10 minutes to segment a unique structure in a CT scan [34]. Recent literature has examined the potential of large-scale *foundation* models in organ segmentation by integrating multiple publicly available datasets corresponding to various tasks [20]. Models that are trained using a wider variety of centers,

acquisition systems, study types, and annotated structures tend to offer better transferability when updated (e.g. fine-tuned) on new tasks and domains. However, since such models are primarily based on public and uncoordinated databases, they may have certain known biases such as long-tail imbalances in the annotated structures or inconsistencies in the annotations [20]. Therefore, identifying efficient and effective learning strategies to adapt these models for new tasks is of high interest in practice.

A popular and widely-adopted approach to adapt a trained model to newly collected data is to fine-tune the whole model on the training images of the novel target task [29]. However, this strategy presents several limitations. First, modern deep-learning models typically have a large number of parameters and require strong hardware requirements for training, which may be unbearable in clinical institutions. Secondly, this approach results in storing a different model for each new domain/task, which might be expensive, especially considering the size of state-of-the-art Transformers-based 3D segmentation networks such as UNETR (#P 555M) [13] or SwinUNETR (#P 371M) [32]. Finally, fine-tuning the entire backbone may lead to suboptimal results, especially when trained on small datasets [21]. An appealing alternative to traditional fine-tuning is parameter-efficient fine-tuning, where only a small subset of parameters are updated during adaptation to new tasks. This family of approaches include, among others, linear probing [23], where only a linear layer staked on top of pre-training features is updated, or adapters [2, 17, 27], which are trainable, compact feed-forward networks that are inserted between the layers of a fixed pre-trained model.

Nevertheless, even though the *pretrain-and-adapt* paradigm is promising, literature on this subject for medical image segmentation is scarce. Furthermore, these works assume that a large set of labeled samples is accessible for the adaptation to the new task. In the medical context, however, since each institute has limited time, budget, and particular clinical purposes, the number of annotated samples available in clinical practice is usually limited. Therefore, it can be very reasonable to assume that adaptation should only be carried out with a few available samples. This motivates the development of new paradigms that allow for resource-efficient adaptation of foundation models in this field.

Our contributions can be summarized as follows:

- We formalize few-shot efficient fine-tuning (FSEFT), a novel and realistic setting for medical image segmentation. We empirically show that, in this setting, standard fine-tuning methods exhibit significant performance drops.
- We introduce a novel parameter-efficient fine-tuning strategy tailored to medical image segmentation, given a handful of labeled samples in the target task. Specifically, we design (a) spatial adapter modules that are more appropriate for dense predictions; and (b) a constrained transductive inference, which leverages task-specific prior knowledge.
- We report comprehensive experiments on a variety of public datasets. The proposed framework approaches full supervision while requiring significantly fewer annotated samples. These results highlight the potential of our framework in practical clinical settings.

2 Related Work

Fine-tuning (FT) has been the most popular approach in the recent years for transferring knowledge across vision tasks [29]. FT involves updating the weights of a pre-trained model by re-training this model with a supervised dataset corresponding to the target task. Particularly in the medical domain, this technique has become integral in a breadth of applications, from radiology [33] to retinal imaging [9,26]. Despite its popularity in medical imaging, FT presents two main limitations. First, these methods are prone to overfitting when the labeled data of the target task is insufficient. Secondly, FT may require substantial computing resources, as it updates all the network parameters, incurring in long adaptation times, more so when using large pre-training models.

Parameter-efficient fine-tuning (PEFT) tackles these limitations, and has recently emerged as an appealing alternative, mostly studied in computer vision and NLP [14]. The problem amounts to adapt a large pre-trained model on new domains/tasks by updating only a small subset of the existing model parameters and/or adding a new, limited set of parameters (a.k.a *adapters*). A simple solution, commonly referred to as Linear Probing (LP) [23, 25], consists of stacking an additional trainable multi-layer perceptron (MLP) layer at the end of the network, whose parameters remain frozen during adaptation. Other strategies include training small auxiliary modules to modify the features extracted by the backbone, such as residual adapters that modify the feature maps [27] or bias parameters [5], batch normalization tuning [24], or side-tuning [36]. More recently, following the popularity of large-scale vision language models, such as CLIP [25], diverse approaches focused on improving their adaption capabilities [11,37]. For instance, [11] proposes CLIP-Adapter, which stacks a small amount of additional learnable bottleneck linear layers to both language and vision branches, while keeping the whole CLIP backbone frozen during adaptation.

Few-shot segmentation (FSS) aims at segmenting novel target classes with just a few labeled samples, where the predominant approach to tackle this problem falls within the meta-learning paradigm. In this scenario, models are trained to learn under few-shot, episodic conditions. In medical image segmentation, the spirit of prototypical networks [30] has been widely adopted, and many mechanisms have been proposed to refine class-wise prototypes. These strategies include, among others, cycle-resemblance attention modules [10], hierarchical attention for time-series consistency [12], iterative refinement through contrastive learning [35] or via local context relationship [31]. However, the meta-learning nature of these approaches hinders the flexible adaptation of foundation models. In particular, their episodic-training procedure implicitly assumes that testing tasks will have a similar structure to the tasks observed during training to reach the best performance [6, 7], i.e., the same number of shots during training and testing. Recent empirical evidence on meta-learning-based foundation models for medical imaging attest to this (see [4], Figure 13). Furthermore, the complexity of the architectural designs associated with many approaches, coupled with the limitation inherited from episodic learning, presents a significant challenge when attempting to utilize existing FSS methods in a PEFT context.

3 Methodology

An overview of our pretraining-adaptation framework is presented in Fig. 1.

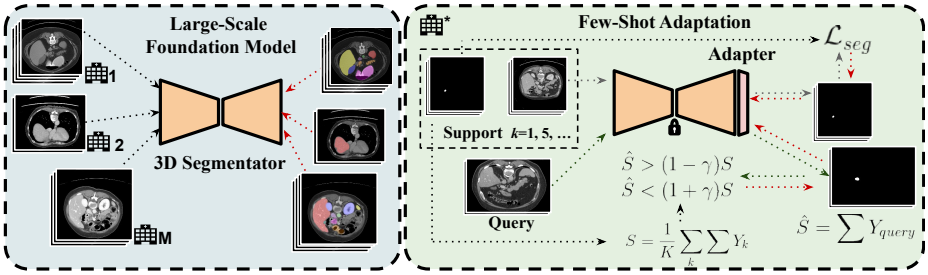

Fig. 1: **Pretrain-and-adapt framework**. We propose to use a foundation model, trained on extensive domains/tasks (see Eq. 1). Then, spatial adapters are trained for institution tuning on a transductive fashion (see Eq. 3).

Foundation model. For training a large-scale foundation model, we consider an assembly of M different datasets, which integrate N different volumes in total. Let $\mathbf{X}_n \in \mathbb{R}^{\Omega_n}$ denotes a medical imaging volume, with Ω_n representing its spatial domain. Each volume is partially annotated at the voxel level, $\mathbf{Y}_n = \{0,1\}^{\Omega_n \times C}$, with C the number of unique categories in the combined dataset. This means that some classes that are considered as foreground in one dataset might be considered as background in another set. Each dataset m presents only partial categories annotated, which are known in the form of a multi-label hot-encoding annotation vector. Thus, each image $\mathbf{X}_n$ is associated with the annotation vector corresponding to its dataset. To simplify the notation, we will denote this vector as $\mathbf{w}_n$, which is directly associated with the dataset to which $\mathbf{X}_n$ belongs. Thus, the training set is composed of the input volumes, their corresponding partial labels, and annotation vectors: $\mathcal{D}_T = \{(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{w}_n)\}_{n=1}^{N}$. Also, let us define a segmentation model, $\theta = \{\theta_f(\cdot), \theta_c(\cdot)\}$, which is composed of a feature extraction neural network, $\theta_f(\cdot)$, and a classification head, $\theta_c(\cdot)$. Thus, the backbone maps each voxel of the input into a spatial feature representation space, $\mathbf{Z}_n = \theta_f(\mathbf{X}_n)$, with $\mathbf{Z}_n \in \mathbb{R}^{\Omega_n \times D}$ and D the number of channels of the output features. Then, the classification head provides a probability distribution $\hat{\mathbf{Y}}_n = \sigma(\theta_c(\mathbf{Z}_n))$, with σ a sigmoid activation. Thus, building the pre-training foundation model θ amounts to using the curated assembled dataset by masked backpropagation of partial labels, and optimizing any segmentation loss function, $\mathcal{L}_{SEG}$, using gradient descent:

$$\min_{\theta_f, \theta_c} \quad \frac{1}{\sum_k \mathbf{w}_{n,k}} \sum_k \mathbf{w}_{n,k} \mathcal{L}_{SEG}(\mathbf{Y}_{n,k}, \hat{\mathbf{Y}}_{n,k}), \quad n = 1, ..., N \tag{1}$$

Few-shot adapters. Inspired by the recent success of adapters in vision-language models and few-shot image classification [11,37], we introduce a PEFT strategy

to adapt a foundation model to new domains in the task of anatomical structure segmentation. Nevertheless, an important difference with respect to existing adapters is that the use of MLP layers is not optimized for dense prediction tasks such as segmentation. Thus, we adopt spatial convolutions in the proposed adapter, which are more suitable for the segmentation problem. Formally, we define a target dataset, $\mathcal{D}^*$, with volumes of an arbitrary study type, X, and a target organ to be segmented, Y. The goal is to adapt the pre-trained foundation model, θ, to the new domain in an efficient way, such as the same backbone, $\theta_f(\cdot)$, is used for different downstream tasks and domains. In addition, we assume that the adaptation should occur using only a few labeled examples (a.k.a the *support set* in the few-shot learning literature [30]), to alleviate the limitation in resources of the target institutions. Thus, a few-shot task includes: (a) A support set of fully-labeled samples, $\mathcal{D}_S = \{(X_k, Y_k)\}_{k=1}^{K}$, with K the total number of support samples (so-called shots), which usually takes small values, i.e. $K = \{1, 5\}$; and (b) a single query (test) volume X for inference. We aim to use this support supervision to train an adapter module, $\phi = \{\phi_f(\cdot), \phi_c(\cdot)\}$, composed of a vision feature extraction based on a few stacked convolutional layers $\phi_f(\cdot)$, and a new classification head, $\phi_c(\cdot)$. The latter yields sigmoid classification scores for both the query and support voxels: $\forall x \in X, \hat{Y}(x) = \sigma(\phi_c(\phi_f(\theta_f(x))))$ and $\forall x \in X_k, \hat{Y}_k(x) = \sigma(\phi_c(\phi_f(\theta_f(x)))), k \in 1, \ldots K$.

Transductive inference leveraging anatomical priors from the support. In image classification, inference is often performed in an *inductive* manner (i.e. one sample at a time). This inductive inference paradigm is common in medical image segmentation, where the task is often seen as a voxel classification. However, segmentation is a *transductive* problem by nature, i.e., one could make joint predictions for all the voxels of the test subject, leveraging available priors on the global structure of the predictions, such the shape of the target organ. Thus, transduction is appealing in our few-shot medical image segmentation setting as the support set could provide such approximate priors. Specifically, we propose to perform inference in a transductive manner, optimizing the cross-entropy loss on the support set while imposing inequality constraints on the size of the target organ in the test subject. Since the volumes are preprocessed to the same resolution, one could estimate an average target-region proportion from the support samples as follows: $S = \frac{1}{K}\sum_k \sum_{x \in \Omega} Y_k(x)$, with Ω denoting the 3D spatial domain. Now, let $\hat{S}$ denotes the predicted size of the target region in the test image, as summation of sigmoid output over the spatial image domain: $\hat{S} = \sum_{x \in \Omega} \hat{Y}(x)$. Thus, we incorporate the following loss during inference, penalizing region proportions that differ from the target by a margin γ:

$$\mathcal{L}_{TI} = \begin{cases} |\hat{S} - (1-\gamma)S|, & \text{if } \hat{S} < (1-\gamma)S \\ |\hat{S} - (1+\gamma)S|, & \text{if } \hat{S} > (1+\gamma)S \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Finally, we train the adapter by gradient steps, integrating the segmentation loss on the support samples and the transductive anatomical constraints in Eq. 2:

$$\min_{\phi_f,\phi_c} \quad \mathcal{L}_{SEG}(Y_k, \hat{Y}_k) + \lambda \mathcal{L}_{TI}(S, \hat{S}_{query}), \quad k = 1, ..., K \tag{3}$$

4 Experiments

Datasets. We use publicly available datasets of partially-labeled CT volumes to build the foundation model and to perform the adaptation experiments. **Foundation model**: a total of 9 datasets with 29 different anatomical structures are assembled. Concretely, BTCV [19], CHAOS [18], LiTS [3], KiTS [15], AbdomenCT-1K [22], AMOS [16], MSD subtasks [1], AbdomenCT-12organ [22] and CT-ORG [28] are gathered to retrieve up to 2022 CT volumes for training. **Adaptation experiments**: We used the TotalSegmentator dataset [34] to evaluate the adaptation of the foundation model, which is composed of 1024 CT volumes with up to 104 anatomical structures, and a wide heterogeneity of scanners and study types. To simulate a real-world use case for adaptation, we retrieved only cases from one of the study types (i.e. CT thorax-abdomen-pelvis) from one institution and selected 9 representative organs present in the foundation model training (i.e. spleen, left kidney, gallbladder, esophagus, liver, pancreas, stomach, duodenum, and aorta). **Pre-processing**: Following previous literature [20], all volumes were standardized and pre-processed to reduce the inter-domain gap. In particular, the orientation of CT volumes was fixed, and isotropic spacing was used to resample the volumes to a voxel size of $1.5 \times 1.5 \times 1.5mm^3$. Finally, the intensity range was clipped to the range [-175, 250], and linearly scaled to [0, 1].

Foundation model pre-training. The partial version of SwinUNETR [32] presented in [20] is used as segmentation architecture. The model is trained on the assembly dataset, by optimizing the Dice loss in Eq. 1 using 3 input patches of size $96 \times 96 \times 96$ per volume in each iteration, with a batch size of 2 volumes, during 120 epochs, and using 4 distributed GPUs. We use AdamW with a base learning rate of $1e^{-4}$, and a warm-up cosine scheduler of 10 epochs. Input patches are augmented through intensity shifts and random rotations of 90 degrees.

Spatial adapter training. In order to adapt the foundation model to new target domains, we freeze the weights of the pre-trained model and remove the classifier head. Over the last feature map, we include as the spatial adapter a decoder block of SwinUNETR. Concretely, the module is composed of three stacked convolutions and one skip connection, with a kernel size of 3, Leaky-ReLU activation, and batch normalization. Under the few-shot learning scenario, the module is trained with k=$\{1, 5, 10\}$ support samples for 100 epochs. We use AdamW and an initial learning rate of 0.5 with cosine decay. The model takes as input 6 patches of size $96 \times 96 \times 96$ per volume in each iteration with a batch size of 1 volume. In the transductive inference (TI) setting, the size constraint over the query sample in Eq. 2 is applied for the last 50 epochs, with $\lambda = 1$.

Evaluation. We benchmark the proposed approach against popular approaches for transfer learning. First, we take as baseline the direct application of the foundation model on the target domain, which we refer to as *Generalization* (i.e. no adaptation). Furthermore, we train from scratch the segmentation model (*Scratch*) using all the available samples for training. As standard fine-tuning approaches, we fine-tune the whole network (*FT*), as well as only the last block (*FT-Last*), where the base learning rate is decreased to $1e^{-4}$. Finally, we include a simple Linear Probe classifier [25] over the features of the pre-trained foundation model. The different approaches are evaluated on five randomly selected query samples retrieved from the subset of each target organ. During testing, the predicted probabilities per voxel are thresholded by a fixed value of 0.5, and the binary mask is post-processed to maintain only the largest connected structure, as in [20]. Last, we use the Dice similarity coefficient (DSC) as evaluation metric.

Ablation experiments. We first assess the effectiveness of the proposed contributions, and motivate our design choices empirically. In particular, Figure 2a depicts the effect of using the proposed spatial adapters compared to an MLP head, which is the dominant approach in previous literature. To better isolate the impact of the adapters module, we evaluate their performance under the standard inductive inference. Results show that by incorporating spatial information the results consistently improve in $[0.7\%, 2.3\%]$ across the different labeled regimes. Second, Figure 2b studies the optimum margin for the size regularizer in the transductive setting Results on three representative structures show that a wide range of γ values offer promising results, above the baseline. However, a lower margin value ($\gamma = 0.05$) might degrade the performance, as the target size is estimated just from a few support samples.

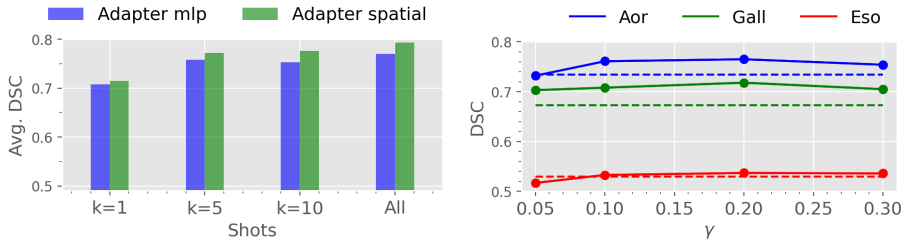

Fig. 2: (a) The role of spatial adapters. (b) Effect of margin γ on the Transductive Inference, with k=10. Dashed lines: no TI. Results using three folds.

Few-shot efficient fine-tuning. Results obtained with the proposed method and relevant baselines are presented in Table 1. ① *Standard fully-supervised regime.* First, results show that the proposed approach brings substantial improvements compared to training the model from scratch, or to the popular LP strategy. Furthermore, it obtains comparable results to fine-tuning the whole model, while updating 300× less parameters. ② *Low-data regime.* As shown in Table 1, fine-tuning the whole model with only a few labeled samples highly deteriorates the generalization performance. While this effect can be mitigated by

only fine-tuning the last layer, this strategy still underperforms recent PEFT alternatives, such as Linear Probe. We can observe that tuning only the proposed spatial adapter head consistently outperforms relevant approaches, with performance gains ranging from 0.7-1.6% in the low-labeled settings compared to the popular LP method [25]. More interestingly, results suggest that, using only one shot and a spatial adapter module, the model generalizes at the same level as training from scratch on the whole dataset. In addition, if we tune only the last convolutional block, which would present the same computational cost as training the spatial adapter, results are consistently worse across each k-shot scenario. This may be explained by the consequent substitution of the rich features learned on a large-scale foundation model. ③ *Constrained transductive inference.* In addition, results show that by incorporating the proposed constrained transductive inference, which leverages task-specific prior knowledge, we approach standard fine-tuning adaptation with the whole dataset in the 10-shot scenario.

Table 1: **Few-shot efficient fine-tuning**. Foundation model adaptation results using different baselines and the proposed adapters on five testing folds.

Setting	Methods	Spl	lKid	Gall	Eso	Liv	Pan	Sto	Duo	Aor	Avg.
	Generalization	0.920	0.891	0.768	0.300	0.950	0.782	0.707	0.363	0.628	0.701
All train (K=∼ 40)	Scratch	0.514	0.896	0.695	0.614	0.902	0.612	0.460	0.552	0.954	0.688
	FT	0.591	0.940	0.654	0.674	0.939	0.853	0.698	0.830	0.926	0.789
	FT-Last	0.954	0.895	0.812	0.423	0.942	0.797	0.784	0.679	0.715	0.777
	Linear Probe [25]	0.948	0.900	0.795	0.422	0.948	0.790	0.773	0.680	0.683	0.771
	Adapter (*Ours*)	0.943	0.904	0.821	0.451	0.948	0.795	0.783	0.669	0.721	**0.781**
10-shot	FT	0.369	0.889	0.249	0.281	0.957	0.454	0.511	0.117	0.917	0.527
	FT-Last	0.960	0.915	0.807	0.425	0.947	0.789	0.723	0.552	0.749	0.763
	Linear Probe [25]	0.942	0.902	0.806	0.452	0.945	0.785	0.786	0.557	0.711	0.765
	Adapter (*Ours*)	0.946	0.900	0.823	0.438	0.945	0.781	0.724	0.704	0.734	0.777
	Adapter + TI (*Ours*)	0.946	0.906	0.821	0.487	0.946	0.785	0.723	0.704	0.735	**0.783**
5-shot	FT	0.553	0.611	0.294	0.586	0.648	0.442	0.164	0.485	0.657	0.493
	FT-Last	0.947	0.712	0.774	0.438	0.952	0.756	0.701	0.619	0.720	0.735
	Linear Probe [25]	0.935	0.887	0.742	0.313	0.960	0.751	0.751	0.525	0.623	0.720
	Adapter (*Ours*)	0.921	0.896	0.822	0.391	0.949	0.752	0.693	0.632	0.680	0.748
	Adapter + TI (*Ours*)	0.928	0.901	0.799	0.442	0.950	0.755	0.712	0.666	0.684	**0.759**
1-shot	FT	0.265	0.255	0.130	0.394	0.519	0.228	0.216	0.162	0.324	0.276
	FT-Last	0.285	0.558	0.366	0.251	0.894	0.585	0.390	0.669	0.394	0.488
	Linear Probe [25]	0.552	0.888	0.671	0.316	0.944	0.488	0.684	0.696	0.679	0.657
	Adapter (*Ours*)	0.549	0.885	0.683	0.351	0.948	0.464	0.703	0.643	0.660	0.654
	Adapter + TI (*Ours*)	0.550	0.888	0.681	0.448	0.947	0.470	0.689	0.631	0.664	**0.663**

#TrainParams: Linear Probe (49) - Adapter/FT-Last (209.6K)

Adapting publicly available pre-trained models. The core idea of the *pretrain-and-adapt* strategy is to allow the data-efficient adaptation of large-scale publicly available models to new scenarios. However, these models for medical volumetric data are scarce. In the following, we present adaptation experiments using the SwinUNETR with the available weights pre-trained on the BTCV dataset from [32] in Table 2. Results suggest the necessity of introducing large foundation models, since direct generalization using dataset-specific models largely drops the performance. When adapting this model using 5 shots, similar trends to the conclusions drawn in the main experiments are observed.

Qualitative evaluation. In the following, we introduce in Figure 3 a qualitative assessment of the performance of the proposed adapter, using the few-shot setting with $k = 5$. The visualizations show the benefits of incorporating

Table 2: **Performance using dataset-specific (BTCV)**, available models.

Setting	Methods	Spl	lKid	Gall	Eso	Liv	Pan	Sto	Aor	Avg.
	Generalization	0.762	0.434	0.398	0.322	0.623	0.458	0.529	0.674	0.524
All train (K=∼ 40)	FT-Last	0.740	0.412	0.419	0.492	0.667	0.510	0.455	0.752	0.555
	Linear Probe [25]	0.576	0.419	0.453	0.327	0.506	0.416	0.458	0.677	0.479
	Adapter (*Ours*)	0.687	0.439	0.522	0.457	0.702	0.532	0.493	0.706	**0.567**
5-shot	FT-Last	0.550	0.405	0.258	0.387	0.722	0.505	0.457	0.732	0.502
	Linear Probe [25]	0.598	0.547	0.078	0.363	0.534	0.352	0.485	0.693	0.456
	Adapter (*Ours*)	0.680	0.496	0.601	0.376	0.585	0.530	0.520	0.676	**0.558**

anatomical constraints regarding organ proportion during adaptation (*first and second rows*). Also, we observe the segmentation improvement of training a small adapter module on top of the backbone for efficient fine-tuning (*second and third rows*).

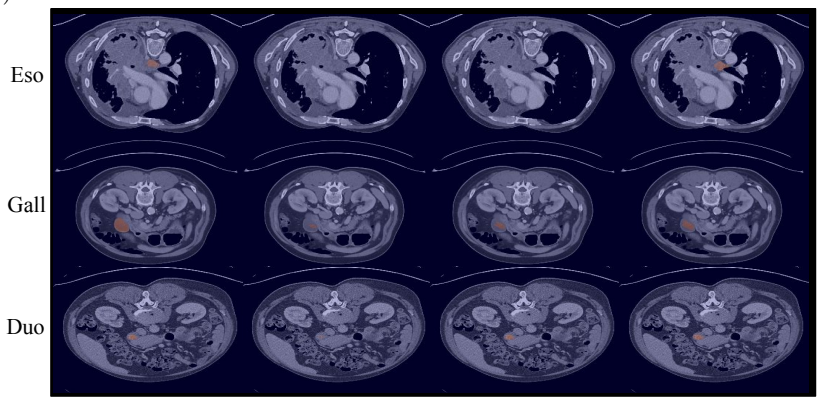

Fig. 3: **Qualitative evaluation**. The axial view of preprocessed CT scans. The annotation/prediction masks of the target organ are in red.

5 Conclusions

The results show the promising performance of our fine-tuning strategy tailored to medical image segmentation, under the *pretrain-and-adapt* paradigm. In particular, we have presented a novel and realistic learning scenario, which accommodates practical clinical settings, i.e., adapting efficiently a large pre-trained model to a new task/domain with a limited number of labeled samples. Our validation demonstrated that standard fine-tuning approaches substantially degrade the performances under this low-data regime. Thus, our results point to the potential of prior-aware transductive inference and spatial adapters in volumetric medical image segmentation.

Acknowledgments

The work of J. Silva-Rodríguez was partially funded by the *Fonds de recherche du Québec (FRQ)* under the Postdoctoral Merit Scholarship for Foreign Students (PBEEE).

References

1. Antonelli, M., *et al.*: The medical segmentation decathlon. Nature Communications **13**, 1–13 (2022)
2. Bapna, A., Arivazhagan, N., Firat, O.: Simple, scalable adaptation for neural machine translation. In: EMNLP (2019)
3. Bilic, P., *et al.*: The liver tumor segmentation benchmark (lits). Medical Image Analysis **84**, 102680 (2023)
4. Butoi, V.I., Ortiz, J.J.G., Ma, T., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Universeg: Universal medical image segmentation. In: ICCV (2023)
5. Cai, H., Gan, C., Zhu, L., Han, S.: Tinytl: Reduce memory, not parameters for efficient on-device learning. In: NeurIPS (2020)
6. Cao, T., Law, M.T., Fidler, S.: A theoretical analysis of the number of shots in few-shot learning. In: ICLR (2020)
7. Chen, W.Y., Liu, Y.C., Kira, Z., Wang, Y.F., Huang, J.B.: A closer look at few-shot classification. In: ICLR (2019)
8. Chen, X., Wang, X., Zhang, K., Fung, K.M., Thai, T.C., Moore, K., Mannel, R.S., Liu, H., Zheng, B., Qiu, Y.: Recent advances and clinical applications of deep learning in medical image analysis. Medical Image Analysis **79**, 4 (2022)
9. De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., et al.: Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature medicine **24**(9), 1342–1350 (2018)
10. Ding, H., Sun, C., Tang, H., Cai, D., Yan, Y.: Few-shot medical image segmentation with cycle-resemblance attention. In: WACV (2023)
11. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. In: ArXiv Preprint (2021), `http://arxiv.org/abs/2110.04544`
12. Guo, S., Xu, L., Feng, C., Xiong, H., Gao, Z., Zhang, H.: Multi-level semantic adaptation for few-shot segmentation on cardiac image sequences. Medical Image Analysis **73**, 102170 (2021)
13. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Daguang, Xu, D.: UNETR: Transformers for 3D medical image segmentation. In: WACV (2022)
14. He, J., Zhou, C., Ma, X., Berg-Kirkpatrick, T., Neubig, G.: Towards a unified view of parameter-efficient transfer learning. In: ICLR (2022)
15. Heller, N., *et al.*: The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the KiTS19 challenge. Medical Image Analysis **67**, 1–16 (2021)
16. Ji, Y., Bai, H., Yang, J., Ge, C., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., Luo, P.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In: NeurIPS (2022)
17. Karimi Mahabadi, R., Henderson, J., Ruder, S.: Compacter: Efficient low-rank hypercomplex adapter layers. In: NeurIPS. pp. 1022–1035 (2021)
18. Kavur, A.E., *et al.*: Chaos challenge - combined (CT-MR) healthy abdominal organ segmentation. Medical Image Analysis **69**, 1–20 (2021)
19. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: MICCAI multi-atlas labeling beyond the cranial vault – workshop and challenge. In: MICCAI Workshop. vol. 5, pp. 1–12. Elsevier B.V. (2015)

20. Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., Landman, B.A., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z.: CLIP-Driven universal model for organ segmentation and tumor detection. In: ICCV (2023)
21. Long, M., Cao, Y., Wang, J., Jordan, M.I.: Learning transferable features with deep adaptation networks. In: ICML (2015)
22. Ma, J., *et al.*: Abdomenct-1k: Is abdominal organ segmentation a solved problem? IEEE Transactions on Pattern Analysis and Machine Intelligence **44**, 6695–6714 (2022)
23. Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., Van Der Maaten, L.: Exploring the limits of weakly supervised pretraining. In: ECCV. pp. 181–196 (2018)
24. Mudrakarta, P.K., Sandler, M., Zhmoginov, A., Howard, A.: K for the price of 1: Parameter-efficient multi-task and transfer learning. In: ICLR (2018)
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICLM. pp. 8748–8763 (2021)
26. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: Understanding transfer learning for medical imaging. In: NeurIPS (2019)
27. Rebuffi, S.A., Bilen, H., Vedaldi, A.: Learning multiple visual domains with residual adapters. In: NeurIPS (2017)
28. Rister, B., Yi, D., Shivakumar, K., Nobashi, T., Rubin, D.L.: Ct-org, a new dataset for multiple organ segmentation in computed tomography. Scientific Data **7**, 1–9 (2020)
29. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2014)
30. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: NeurIPS (2017)
31. Tang, H., Liu, X., Sun, S., Yan, X., Xie, X.: Recurrent mask refinement for few-shot medical image segmentation. In: ICCV (2021)
32. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3D medical image analysis. In: CVPR (2021)
33. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: CVPR (2017)
34. Wasserthal, J., Meyer, M., Breit, H.C., Cyriac, J., Yang, S., Segeroth, M.: Totalsegmentator: robust segmentation of 104 anatomical structures in ct images. Radiology: Artificial Intelligence **5** (2023)
35. Wu, H., Xiao, F., Liang, C.: Dual contrastive learning with anatomical auxiliary supervision for few-shot medical image segmentation. In: ECCV (2022)
36. Zhang, J.O., Sax, A., Zamir, A., Guibas, L., Malik, J.: Side-tuning: A baseline for network adaptation via additive side networks. In: ECCV (2019)
37. Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free clip-adapter for better vision-language modeling. In: ECCV (2022)