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

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

ETS Montréal, Canada

Take-Home Message

Proposal:

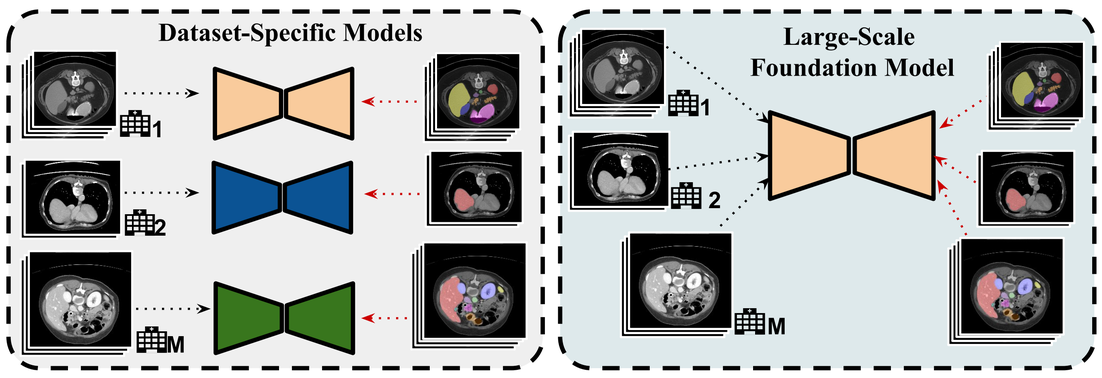
- Few-shot parameter-efficient fine-tuning (FSEFT), a novel and realistic setting for adapting volumetric foundation models on clinical scenarios.
- Using spatial adapter modules, tailored for medical image segmentation, and a transductive inference to leverage priors during adaptation.

Results:

- In the few-shot regime, standard finetuning exhibits performance drops.
- The potential of foundation models: only 5-shots and classifier adapters outperform training from scratch on the whole dataset.
- Our solution provides better performance while updating 300× less parameters.

Introduction

- **Context:** Foundation models for volumetric medical images [1].
  - Capture rich features pre-trained on large, heterogeneous sources.
  - Promising transferability to downstream datasets/tasks.



- **Motivation:** The real-world clinical scenario:
  - Scarce data and annotation resources.
  - Limited computation power.
- **Background:** Standard adaptation methods such as fine-tuning (FT) fail in this scenario.

Setting	Methods	Avg. DSC
10-shot	FT	0.527
	FT-last	0.763
	Linear Probe [2]	0.777

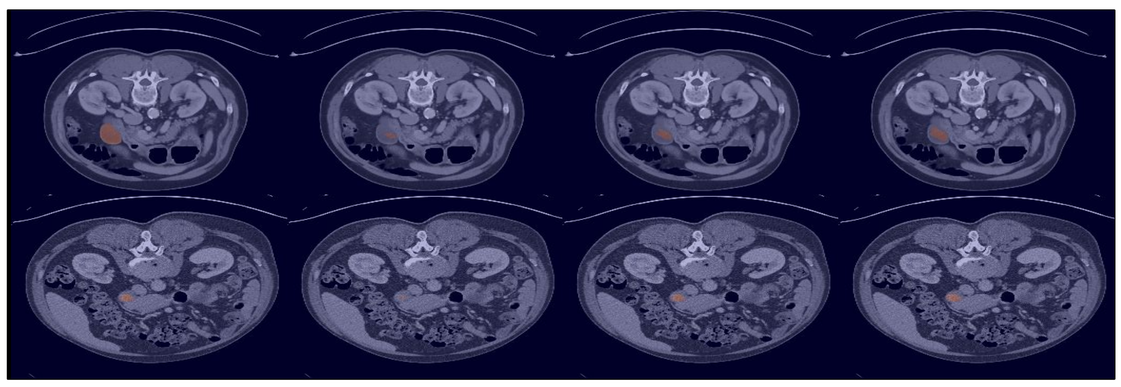
- **Contribution:**
  - We formalize few-shot efficient fine-tuning (FSEFT), a novel and realistic setting for medical image segmentation.
  - We design spatial adapter modules that are more appropriate for dense predictions.
  - We introduce a constrained transductive inference, which leverages task-specific prior knowledge.
  - The proposed framework approaches full supervision while requiring significantly fewer annotated samples

Results

(1) Quantitative performance:

Methods	Avg. DSC			
	1-shot	5-shot	10-shot	All
Scratch	-	-	-	0.688
FT	0.276	0.493	0.527	**0.789**
FT-last	0.488	0.735	0.763	0.777
Linear Probe [2]	0.657	0.720	0.765	0.771
Adapter (Ours)	0.654	0.748	0.777	0.781
Adapter + TI (Ours)	**0.663**	**0.759**	**0.783**	-

(2) Qualitative results:



Method

- **Foundation Model:** We use an assembly dataset of volumetric data $\mathcal{D}_T = \{(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{w}_n)\}_{n=1}^{N}$, composed by pre-processed volumes, pixel-level partial annotations, and task masks, respectively. Also, let us define a segmentation model, $\theta = \{\theta_f(\cdot), \theta_c(\cdot)\}$, which is composed of a feature extraction neural network and a classification head, such that, $\hat{\mathbf{Y}}_n = \sigma(\theta(\mathbf{X}_n))$. The pre-training of a foundation model consists of optimizing any segmentation loss for each k task under:

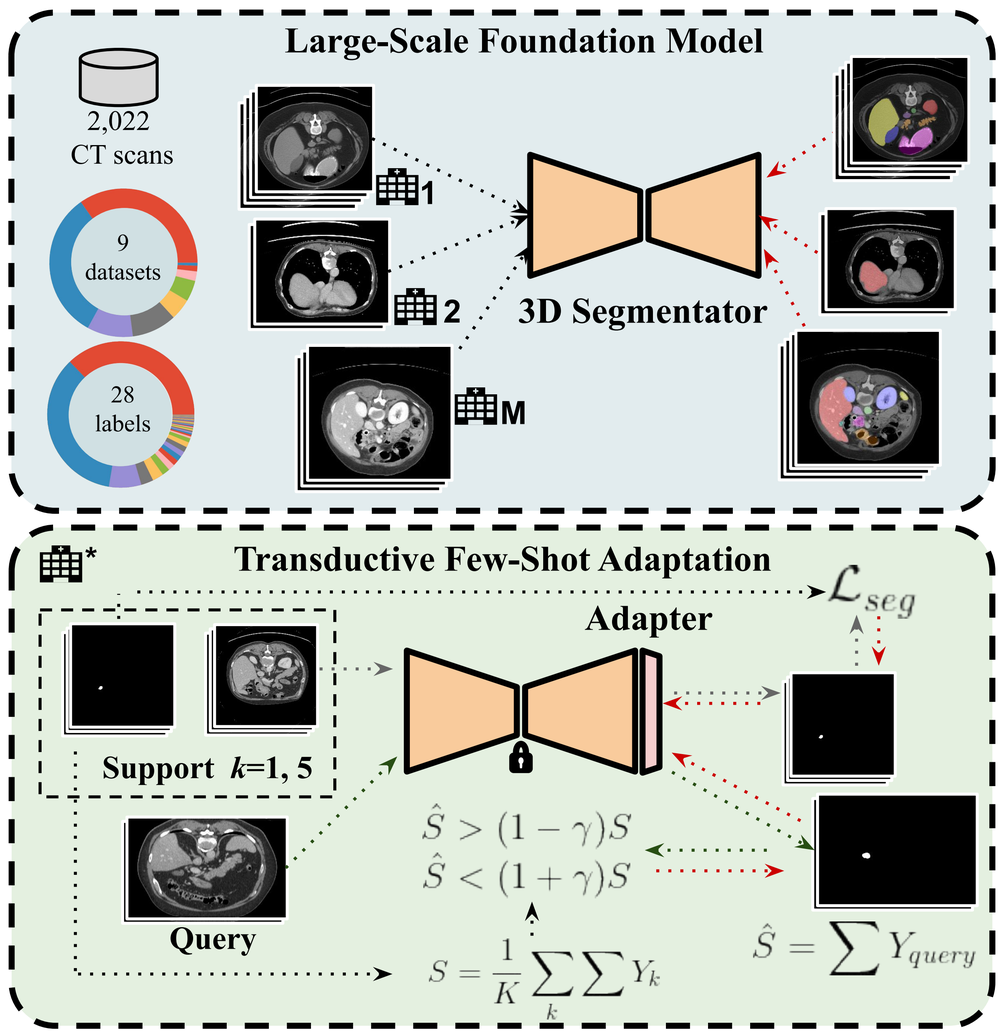
$$\min_{\theta_f, \theta_c} \quad \frac{1}{\sum_k \mathbf{w}_{n,k}} \sum_k \mathbf{w}_{n,k} \mathcal{L}_{SEG}(\mathbf{Y}_{n,k}, \hat{\mathbf{Y}}_{n,k}), \quad n = 1, ..., N \tag{1}$$

- **Parameter-Efficient Few-shot Adapters:** During adaptation, we replace the classification head with an adapter module, ϕ. This module produces voxel-level sigmoid scores for the query sample and few support, annotated examples: $\forall x \in X, \hat{Y}(x) = \sigma(\phi(\theta_f(x)))$ and $\forall x \in X_k, \hat{Y}_k(x) = \sigma(\phi(\theta_f(x))), k \in 1, \dots K$.
- **Transductive Inference:** Anatomical priors - in the form of organ size, $S = \frac{1}{K} \sum_k \sum_{x \in \Omega} Y_k(x)$, are incorporated during inference to enhance the consistency on the query sample prediction.

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

Adaptation stage:

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



- **Repository:** https://github.com/jusiro/fewshot-finetuning

References

[1] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A. Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. CLIP-Driven universal model for organ segmentation and tumor detection. In *ICCV*, 2023.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICLM*, 2021.

The work of J. Silva-Rodríguez was partially funded by the FRQ under the PBEEE merit scholarship.