ÉTS
Le génie pour l'industrie

MICCAI 2023 Vancouver CANADA

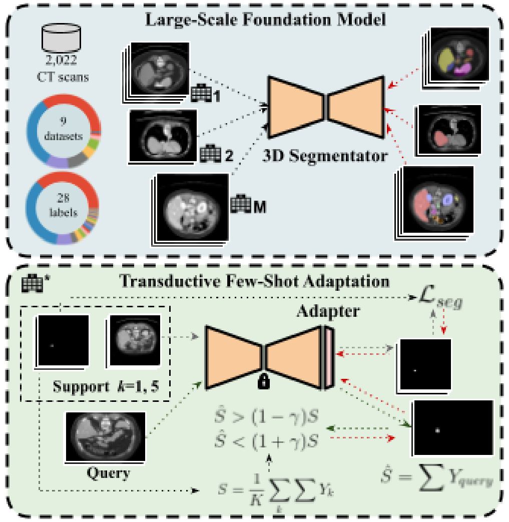
MedAGI

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

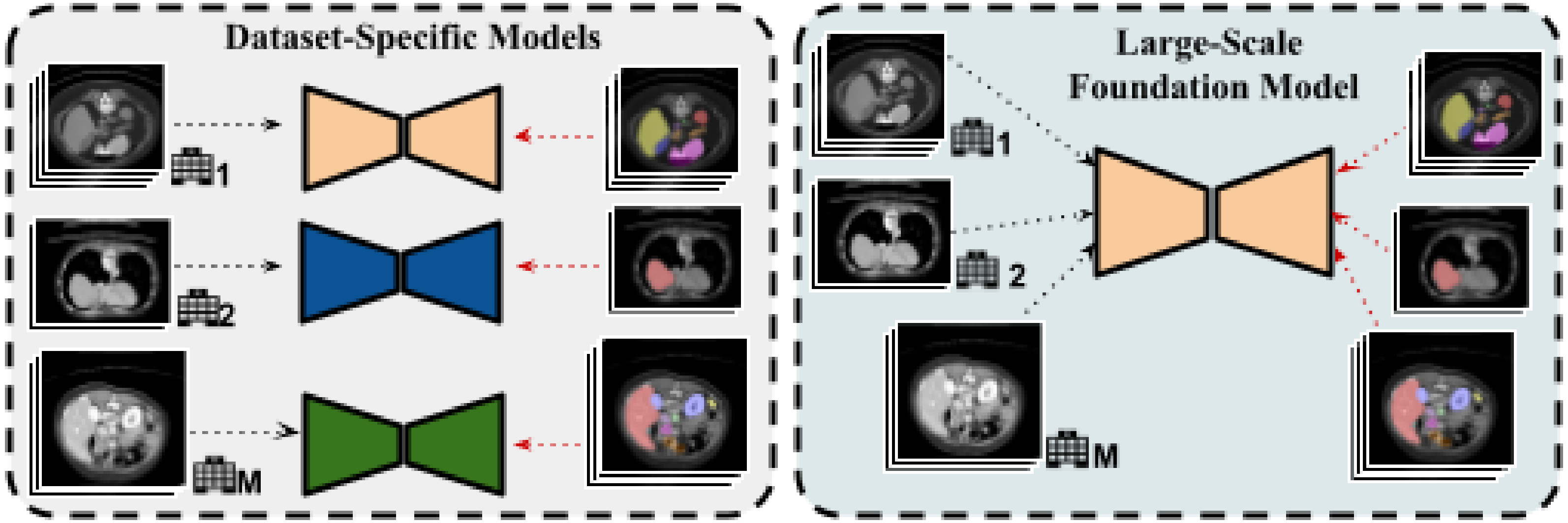
https://github.com/jusiro/fewshot-finetuning



Towards foundation models for volumetric segmentation

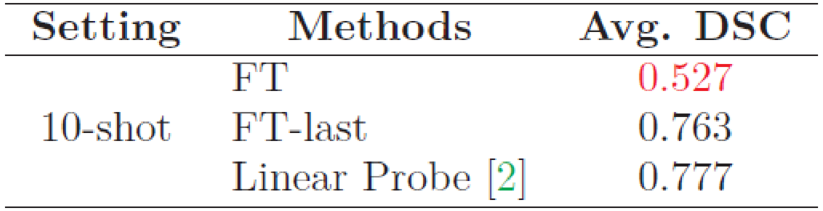
- **Foundation models** are in their **early stages** for medical volume segmentation.

- Some works have already shown their **generalization/transferability potential**: CLIP-driven Universal Model (*Liu et al.* 23), Uniseg (*Ye et al.* 23).



Pretrain-and-Adapt: real world requirements

- An experienced clinician requires an average of **10 minutes to segment an unique structure in a CT scan** (*Wasserthal et al* 23).
- Current deep-learning models are huge (#P 555M), and so are CT volumes. **Clinical institutions have limited computational resources.**
- Current **adaptation strategies are not prepared for this setting**.

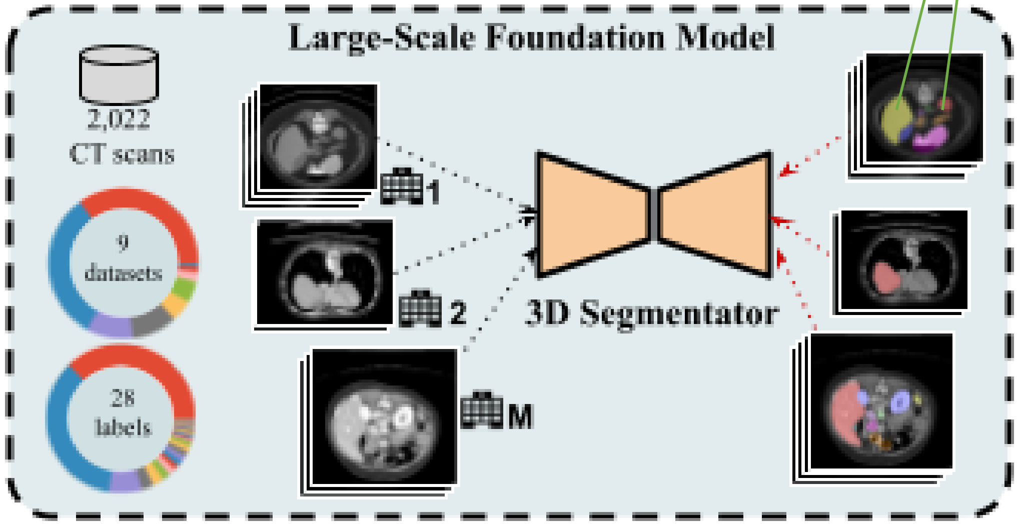


Setting	Methods	Avg. DSC
10-shot	FT	0.527
	FT-last	0.763
	Linear Probe [2]	0.777

Foundation model pre-training

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

masked backprop. on the annotated tasks per sample



- **9 Datasets (2022 CT scans):** BTCV, CHAOS, LiTS, KiTS, AbdomenCT-1K, AMOS, MSD, AbdomenCT-12 organs, CT-org.

- **29 Tasks**: spleen, rkidney, lkidney, gall, esophagus, liver, stomach, aorta, postcava, psv, pancreas, radrenal, ladrenal, duodenum, bladder, prostate, uterous, liver tumor, kidney tumor, kidney cyst, celiac truck, lung, bone, brain, lung tumor, pancreas tumor, colon tumor…

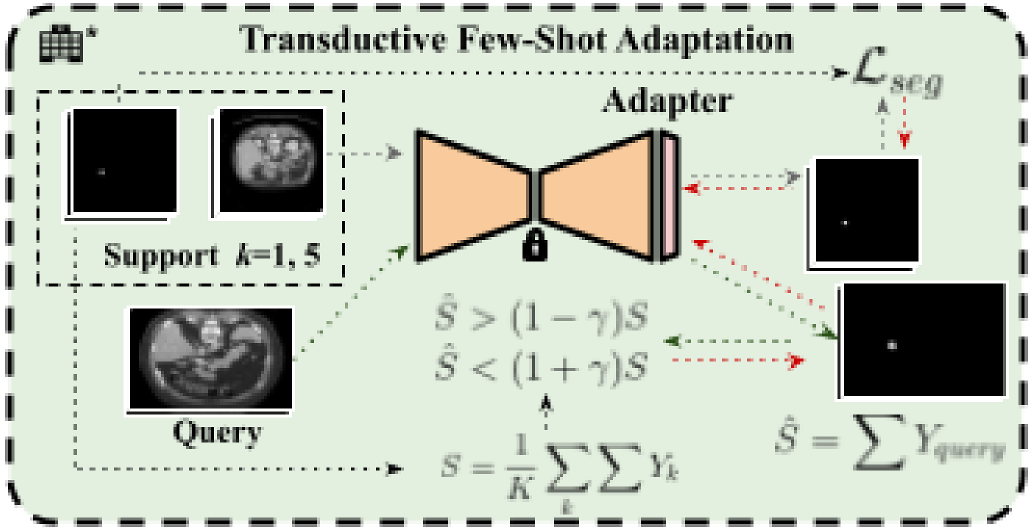
- **Implementation**: SwinUNETR (*Tang et al.* 21) with sigmoid outputs and DICE Loss.

Parameter-Efficient Few-Shot Adapters

- **Efficient Transfer Learning:** using the **frozen pre-trained model**, we **replace the classification head**, and add a new one (i.e. adapter), **including convolution blocks**.

- **Few-shot adaptation stage**: taining on ***k* support examples**, and testing on **one query sample**.

$$\min_{\phi} \quad \mathcal{L}_{SEG}(Y_k, \hat{Y}_k), \quad k = 1, \dots, K$$



- **1 unseen dataset:** TotalSegmentator.

- **9 Tasks**: spleen, left kidney, gallbladder, esophagus, liver, pancreas, stomach, duodenum, aorta.

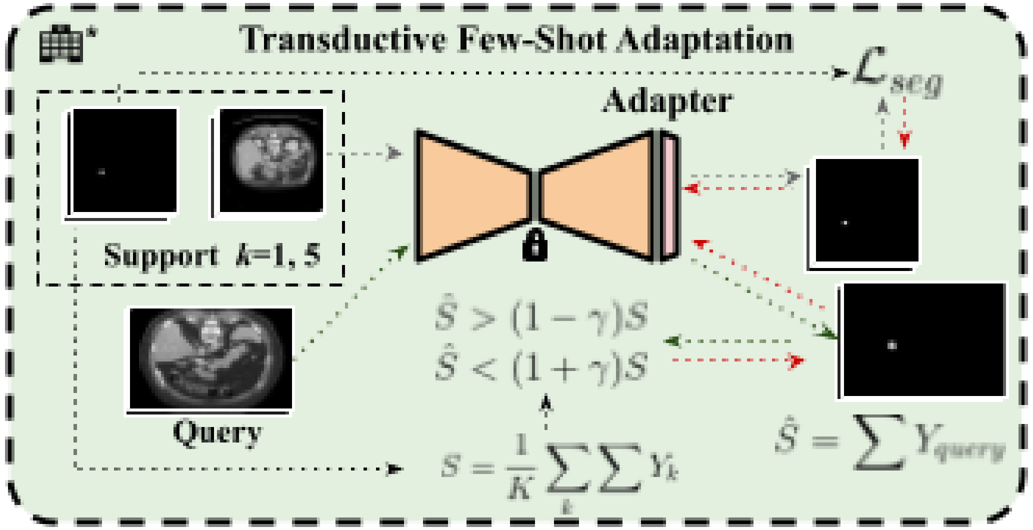
- **Implementation**: The spatial adapter contains one randomly initialized convolution block from SwinUNETR decoder.

- Adaptation is performen for each organ individually.

Incorporating priors during adaptation

- **Constraining to proper priors on the query sample**: we can estimate the organ size (S) on the support set and enhance the adaptation stage in a **transductive way**.

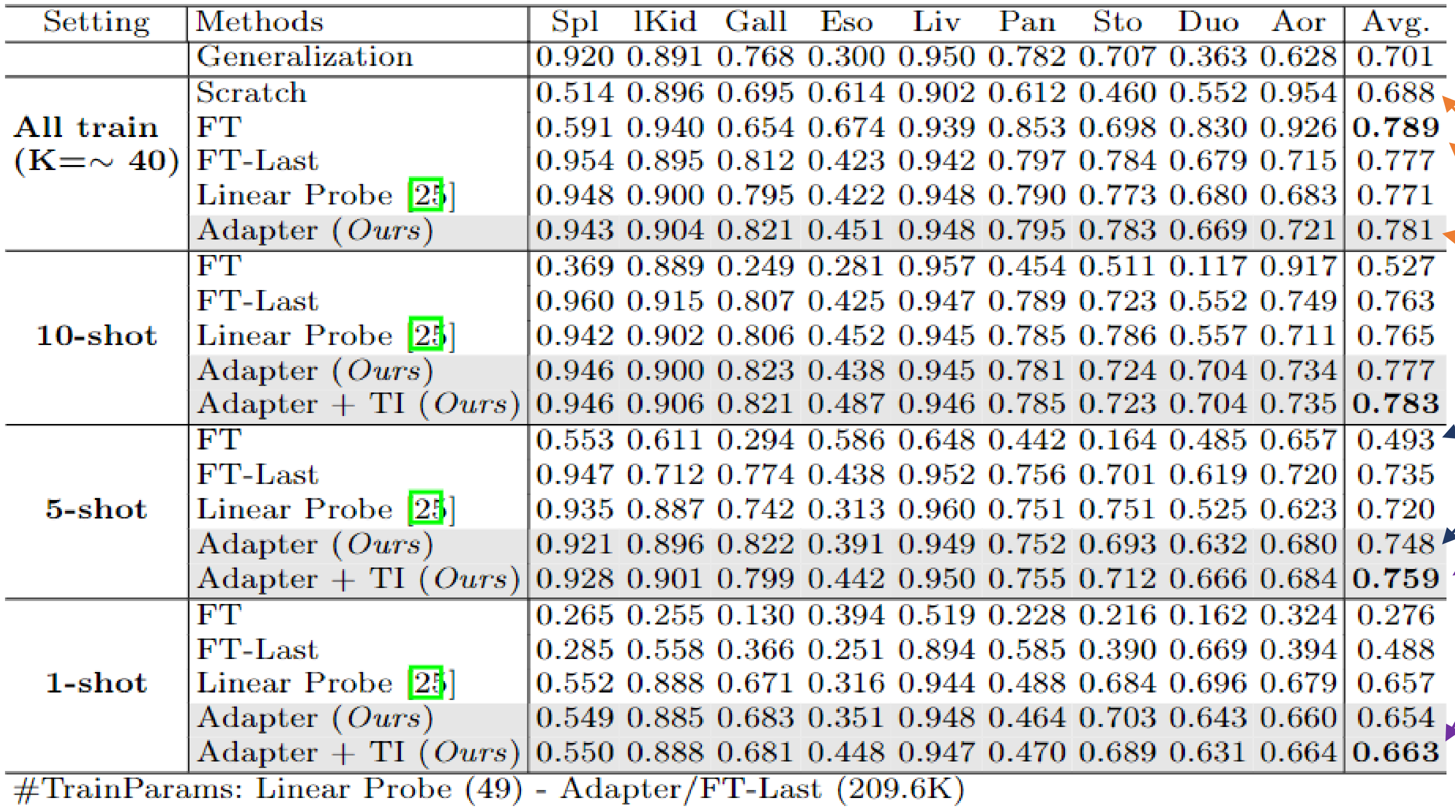
$$\mathcal{L}_{TI} = \begin{cases} |\hat{S} - (1-\gamma)S|, & \text{if } \hat{S} < (1-\gamma)S \\ |\hat{S} - (1+\gamma)S|, & \text{if } \hat{S} > (1+\gamma)S \\ 0, & \text{otherwise} \end{cases}$$



- **Transductive adaptation stage**:

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

Results



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)

1. Standard fully-supervised regime.

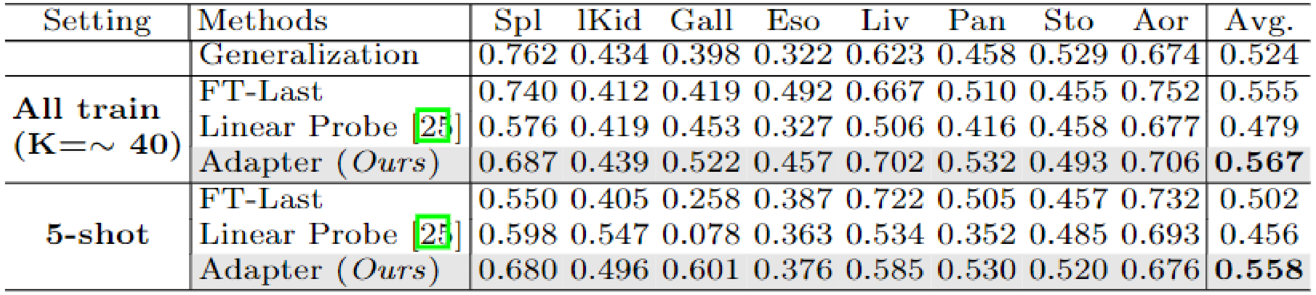
2. Low-data regime.

3. Incorporate priors.

Results

- Are **current available models** prepared for this setting?

Using pre-trained weights from SwinUNETR pre-trained on BTCV(*Tang et al.* 21)



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**

- Our **pre-trained weights are publicly available**:

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

Take-home messages

- In the **clinical scenario**, the **adaptation** of **foundation models** should require low data (**few-shots**) and **limited computational resources**.

- In this scenario, **standard fine-tuning exhibits performance drops**.

- **Few-shot parameter-efficient fine-tuning (FSEFT)**: a novel and realistic setting for adapting volumetric foundation models on clinical scenarios .

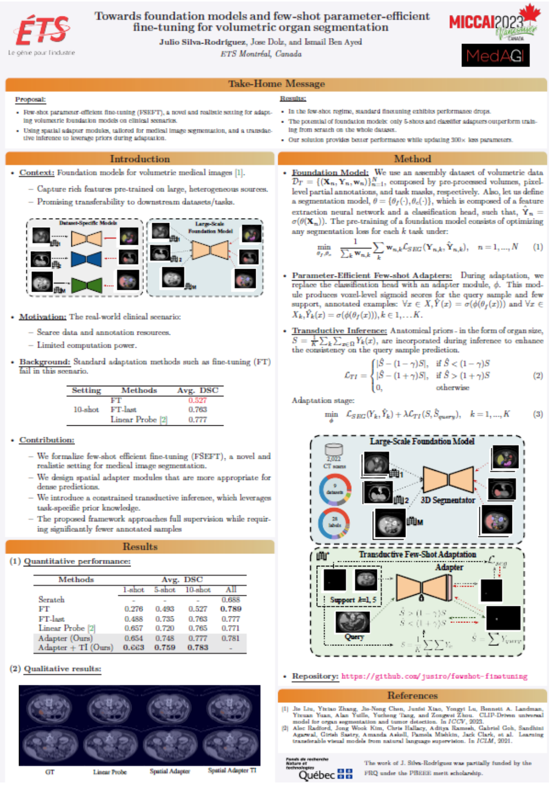
- You can design ad-hoc **adapters** and incorporate **priors** during the adaptation.

- **Potential**: only **5-shots** outperform training from scratch on the whole dataset – and **300x less parameters**.

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



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:

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

- **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 $\forall x \in X_k, \hat{Y}_k(x) = \sigma(\phi(\theta_f(x))), k \in 1, \ldots K$.
- **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.

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

Adaptation stage:

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

- **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.