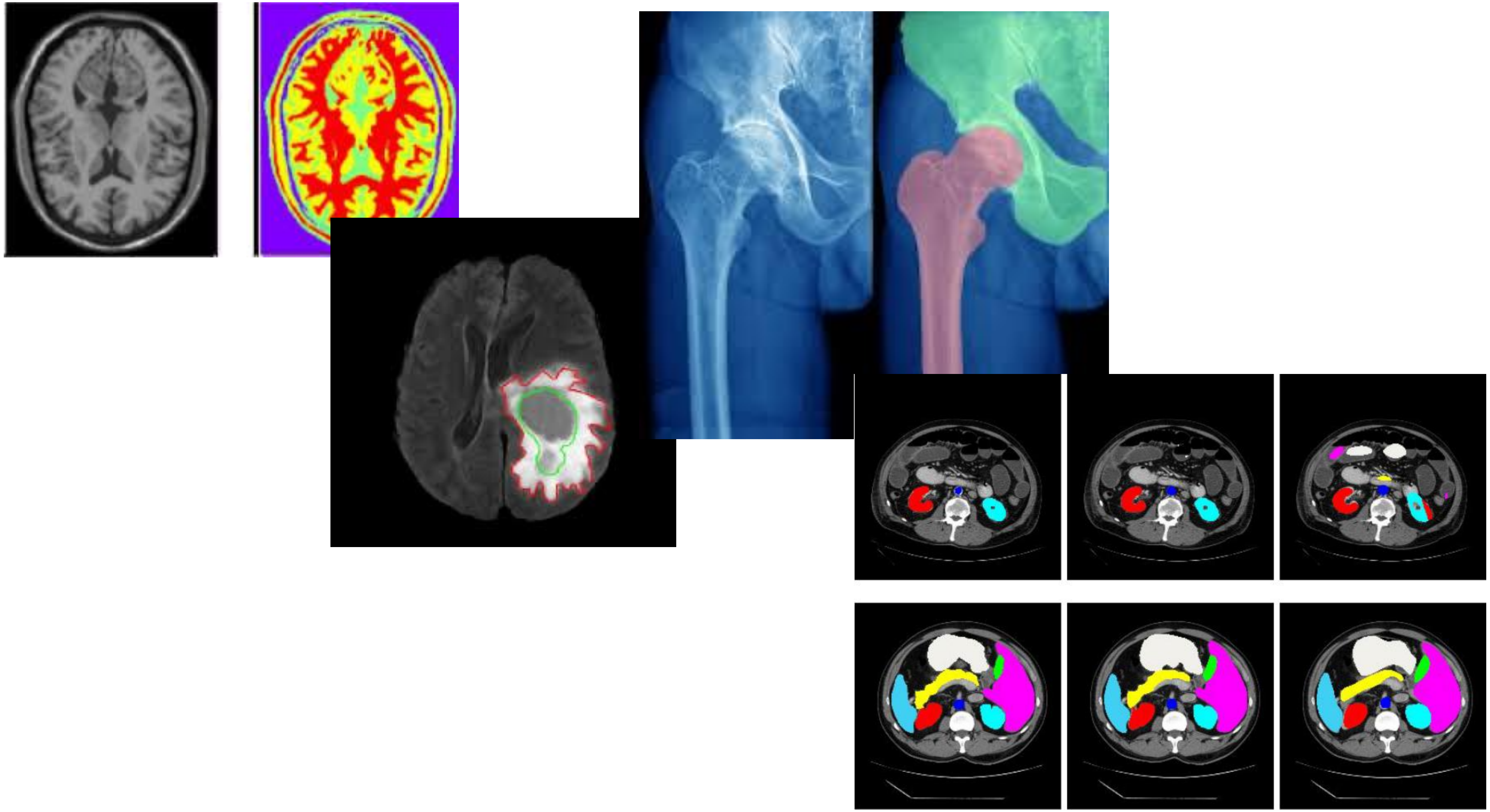
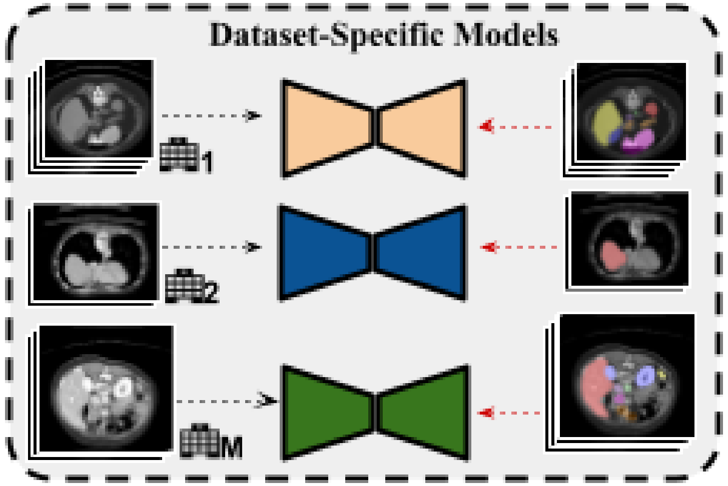
Foundation models for medical image segmentation

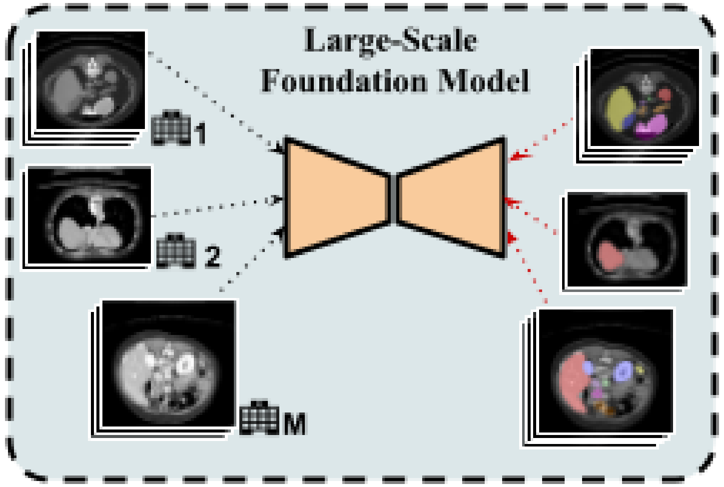


Foundation models for medical image segmentation



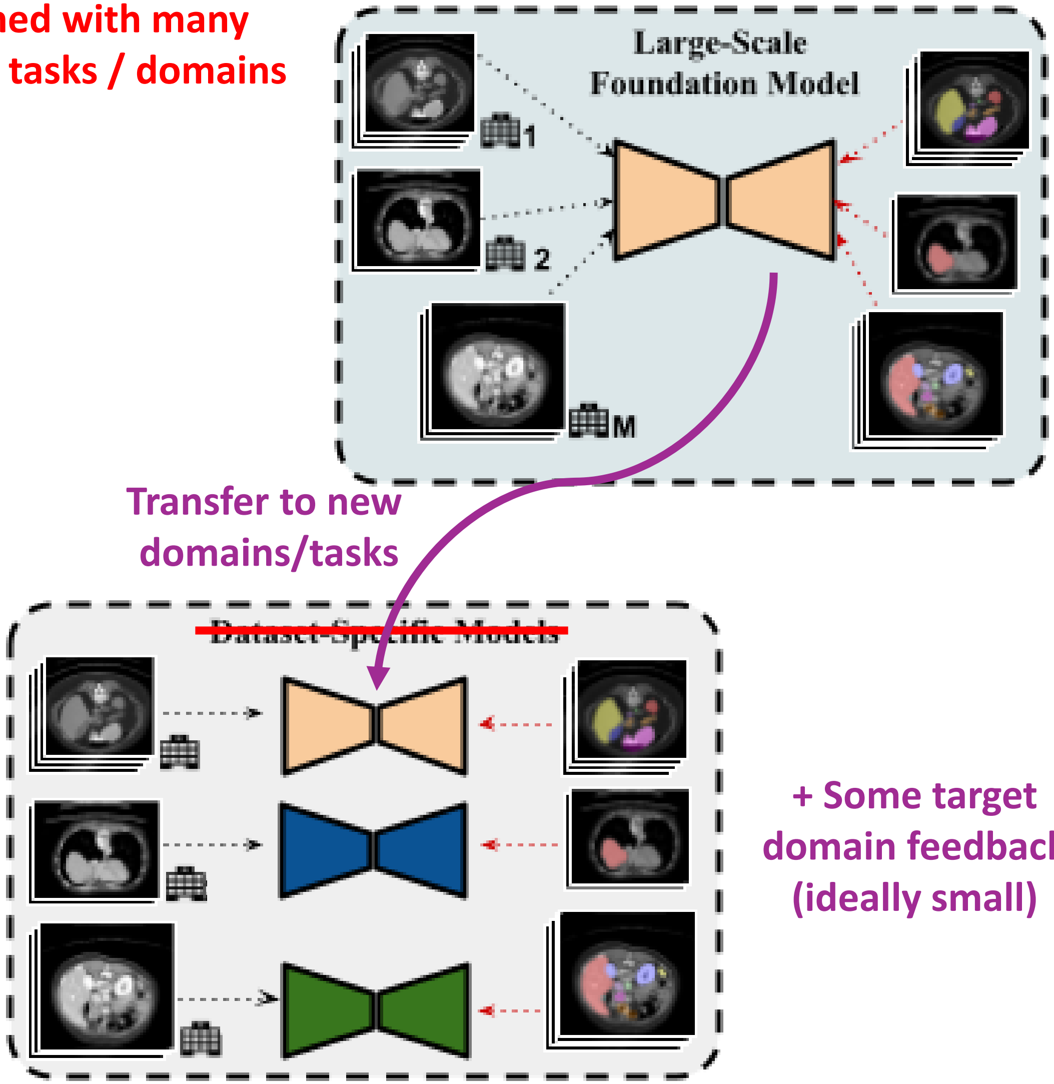
Foundation models for medical image segmentation

Trained with many data / tasks / domains



Foundation models for medical image segmentation

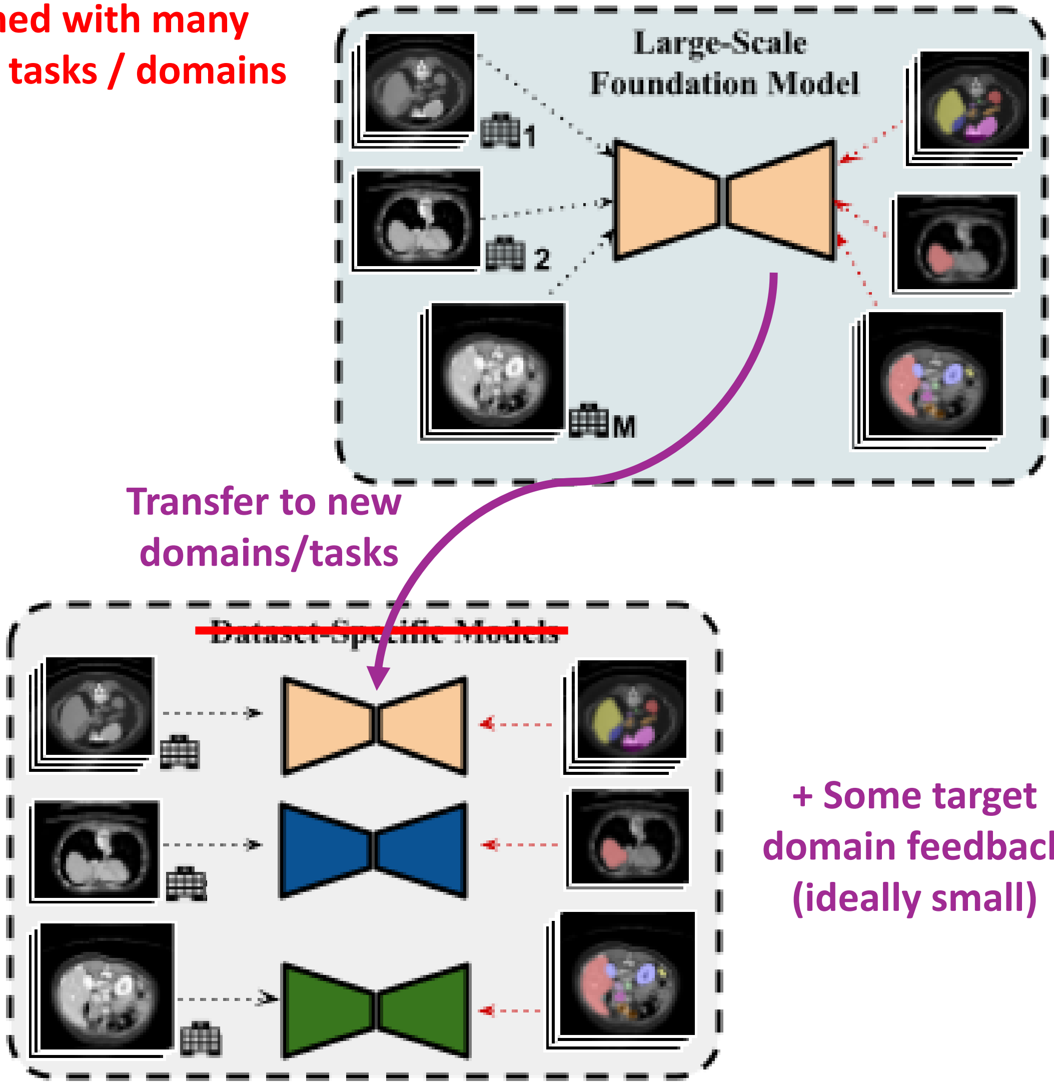
Trained with many data / tasks / domains



+ Some target domain feedback (ideally small)

Foundation models for medical image segmentation

Trained with many data / tasks / domains



Transfer to new domains/tasks

+ Some target domain feedback (ideally small)

UniverSeg

UniSeg

MedSAM

CLIP-Driven Universal Model

SuPreM

MultiTalent

MedSAM-3D

Hermes

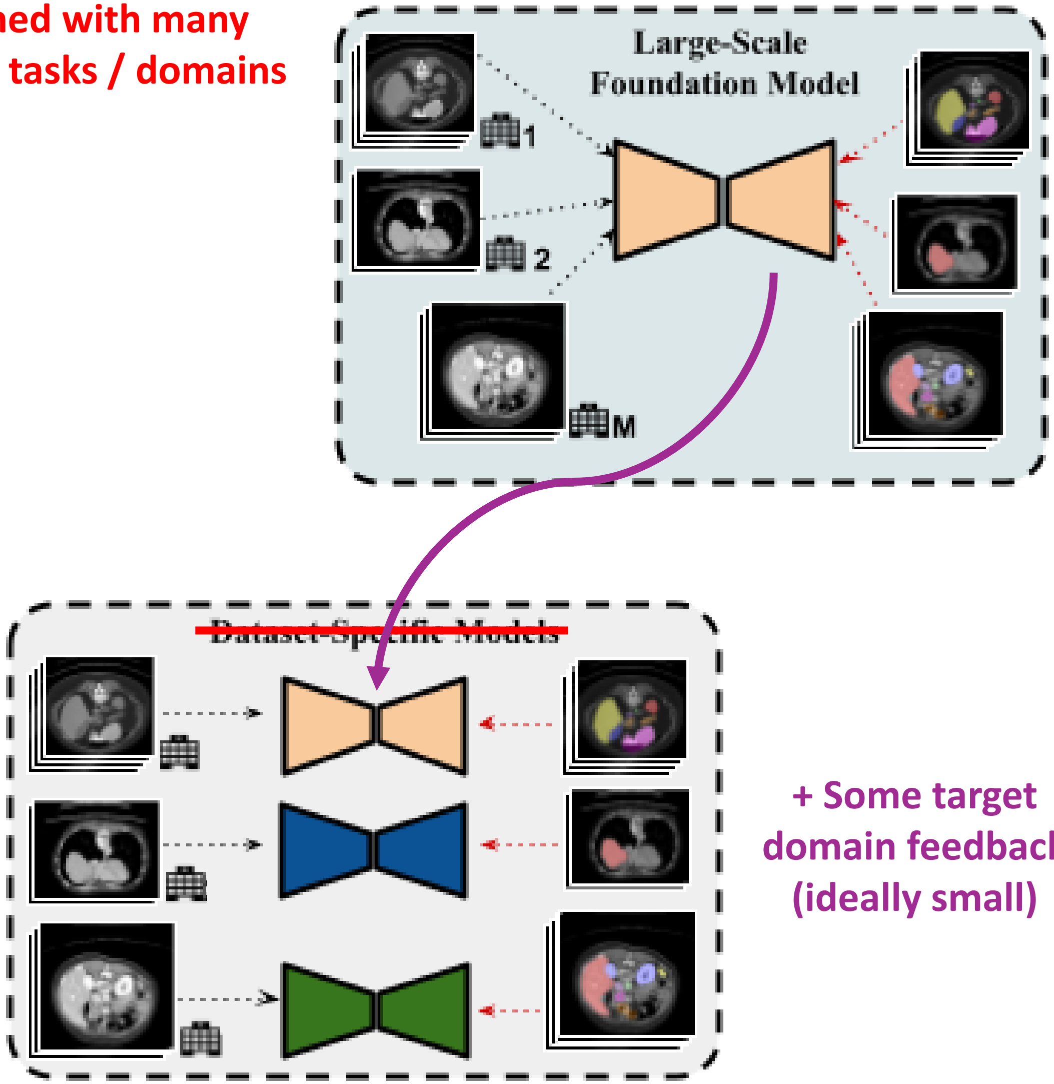
SAM

MA-SAM

2023-2024-...

Foundation models for medical image segmentation

Trained with many data / tasks / domains



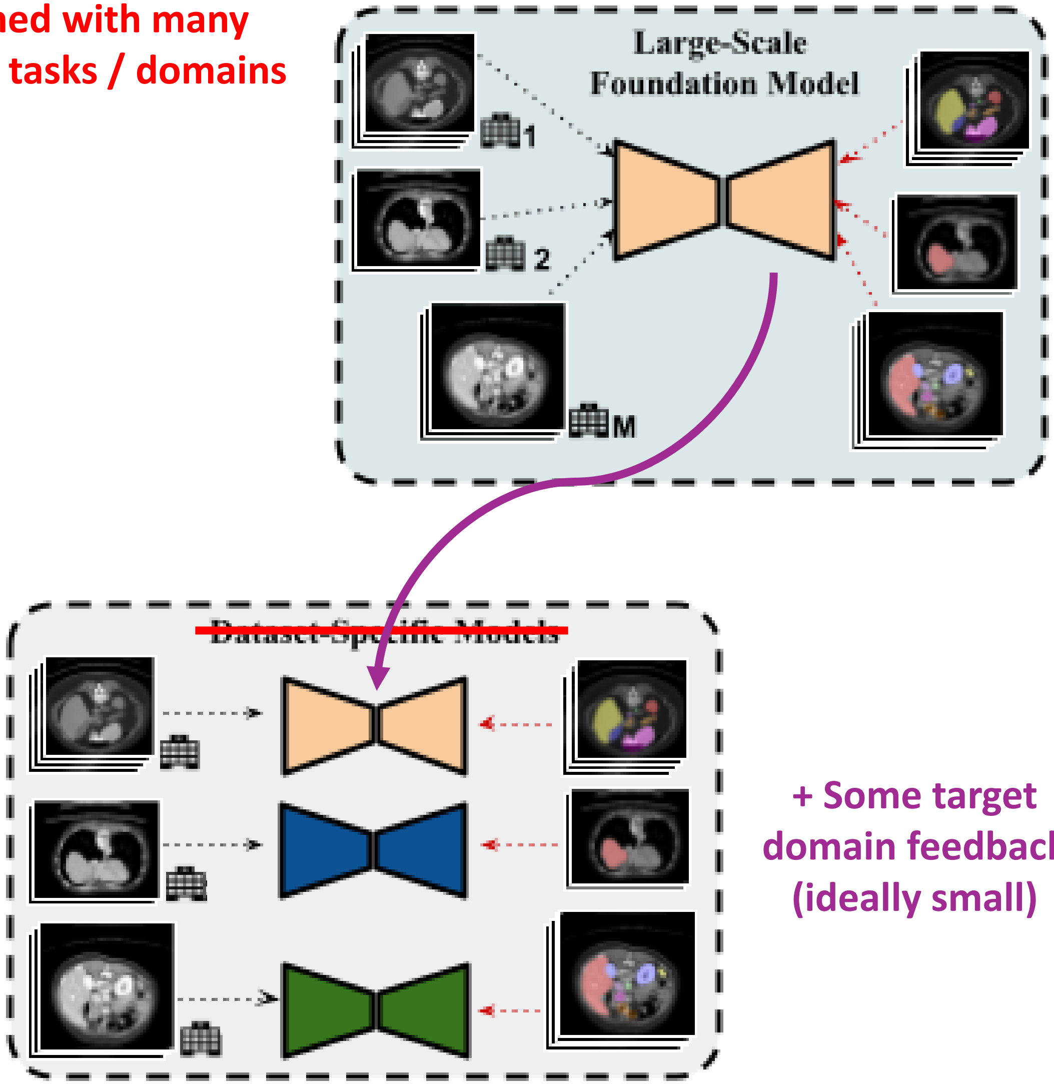
+ Some target domain feedback (ideally small)

Organizing the mess!

1. Types of foundation models: a data perspective.
    A. Generalist vs. Specialized
    B. 2D *vs*. 3D
    C. Multimodal *vs*. Unimodal
2. Learning/Usage Objectives
    A. Zero-shot / Transfer Learning
    B. In-Context Learning
    C. Interactive Models ("SAM")
3. Zero-shot / Adaptation-oriented (3D data)
    A. How to pre-train?
    B. How useful are foundation models? Limitations on the adaptation stage
    C. Few-shot Parameter-Efficient Fine-tuning

Foundation models for medical image segmentation

Trained with many data / tasks / domains



+ Some target domain feedback (ideally small)

Organizing the mess!

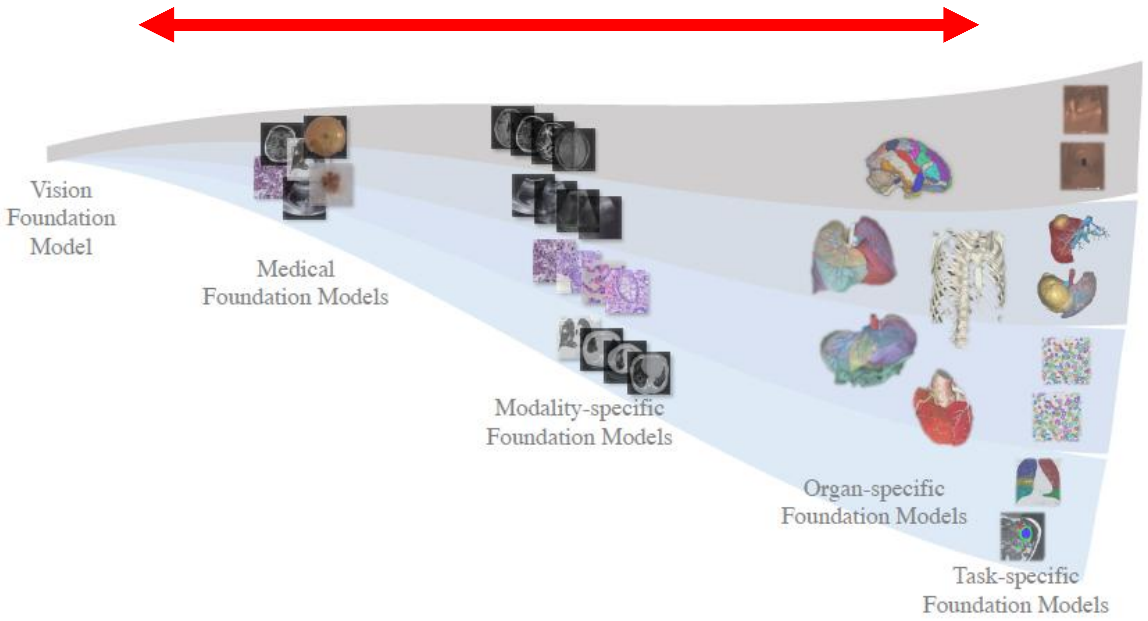
1. **Types of foundation models: a data perspective.**
   - A. **Generalist vs. Specialized**
   - B. **2D *vs*. 3D**
   - C. **Multimodal *vs*. Unimodal**
2. Learning/Usage Objectives
   - A. Zero-shot / Transfer Learning
   - B. In-Context Learning
   - C. Interactive Models (“SAM”)
3. Zero-shot / Adaptation-oriented (3D data)
   - A. How to pre-train?
   - B. How useful are foundation models? Limitations on the adaptation stage
   - C. Few-shot Parameter-Efficient Fine-tuning

Types of foundation models: a data perspective.

Generalist *vs*. Specialized (pre-training)

+ Data Available

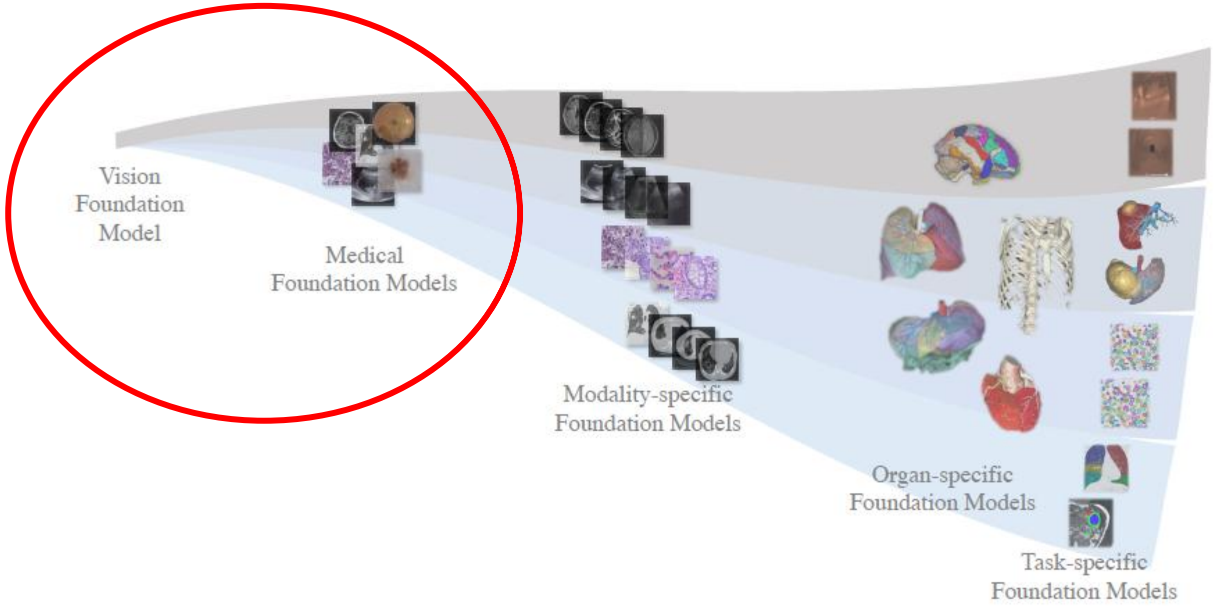
- Data Available



Huang et *al*. On The Challenges And Perspectives of Foundation Models For Medical Image Analysis. MedIA'24.

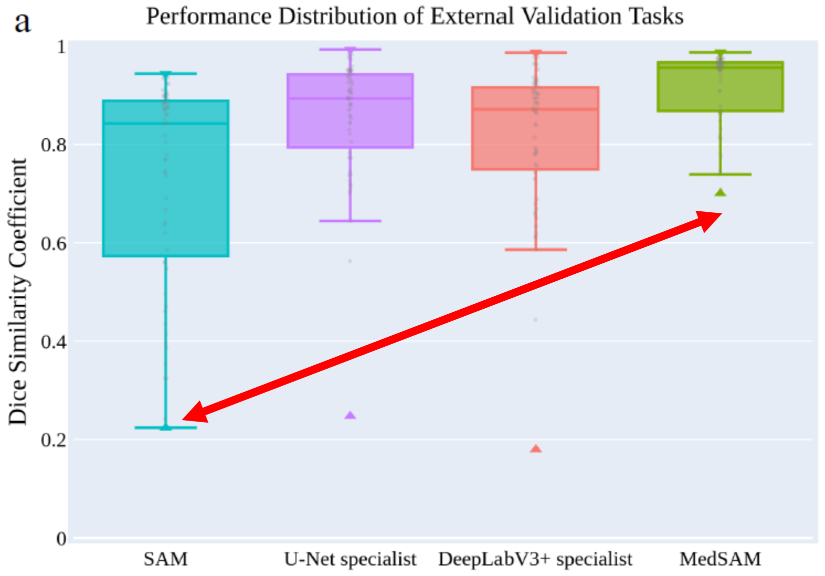
Types of foundation models: a data perspective.

Generalist *vs.* Specialized (pre-training)



Huang et *al*. On The Challenges And Perspectives of Foundation Models For Medical Image Analysis. MedIA'24.

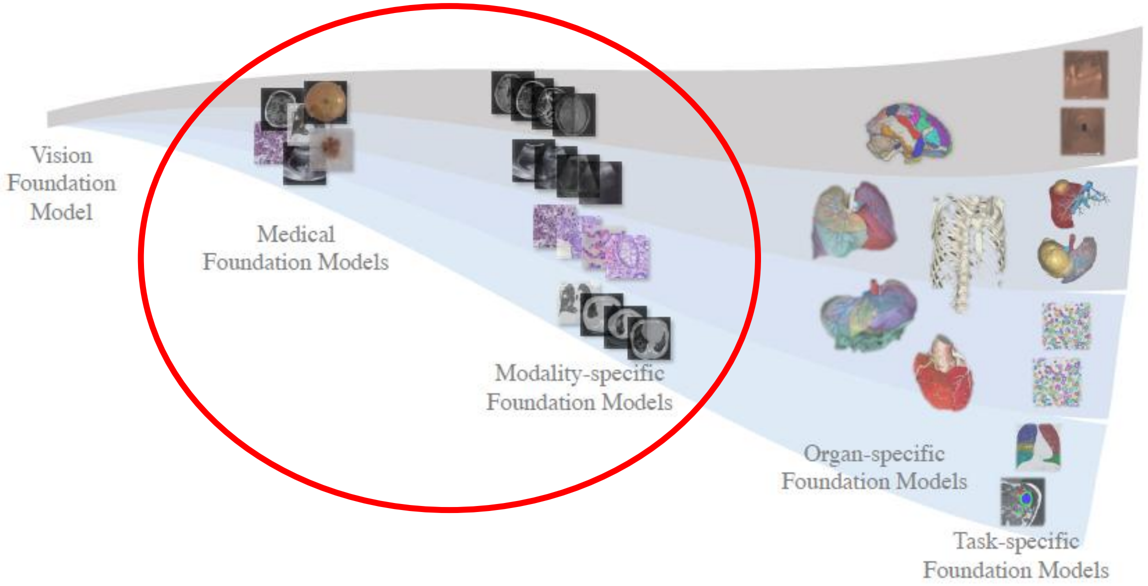
→ **Medical better than General** (natural image)



Ma et al. Segment Anything in Medical Images. Nat.Com.'24

Types of foundation models: a data perspective.

Generalist *vs.* Specialized (pre-training)

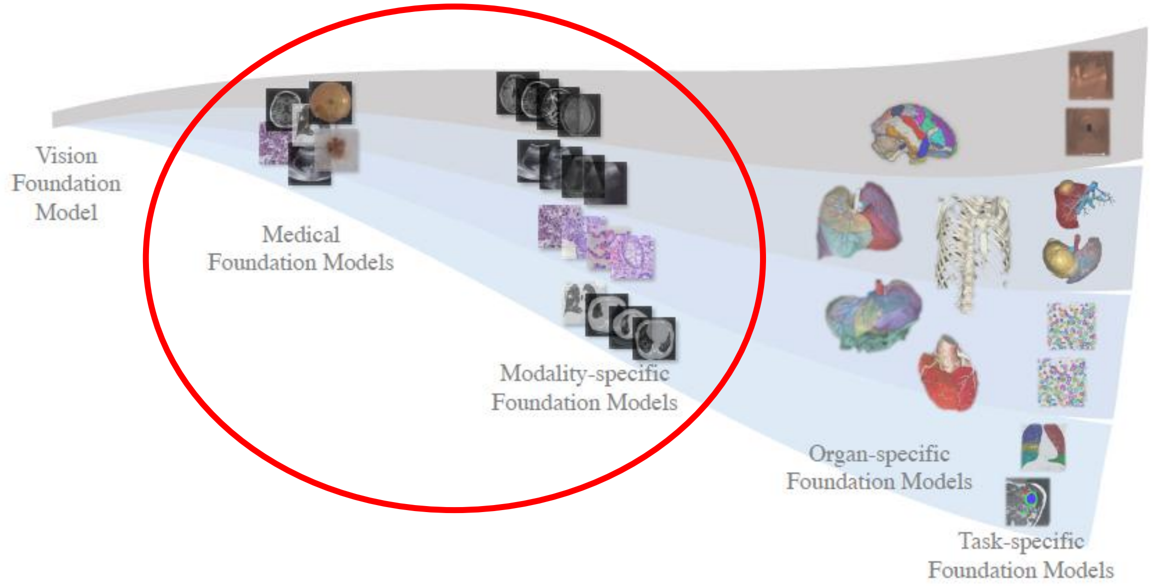


Huang et *al*. On The Challenges And Perspectives of Foundation Models For Medical Image Analysis. MedIA'24.

→ **Modality better than Medical ?**
(scarce empirical studies for segmentation)

Types of foundation models: a data perspective.

Generalist *vs.* Specialized (pre-training)



Huang et *al*. On The Challenges And Perspectives of Foundation Models For Medical Image Analysis. MedIA’24.

→ **Modality better than Medical ?**

(scarce empirical studies for segmentation)

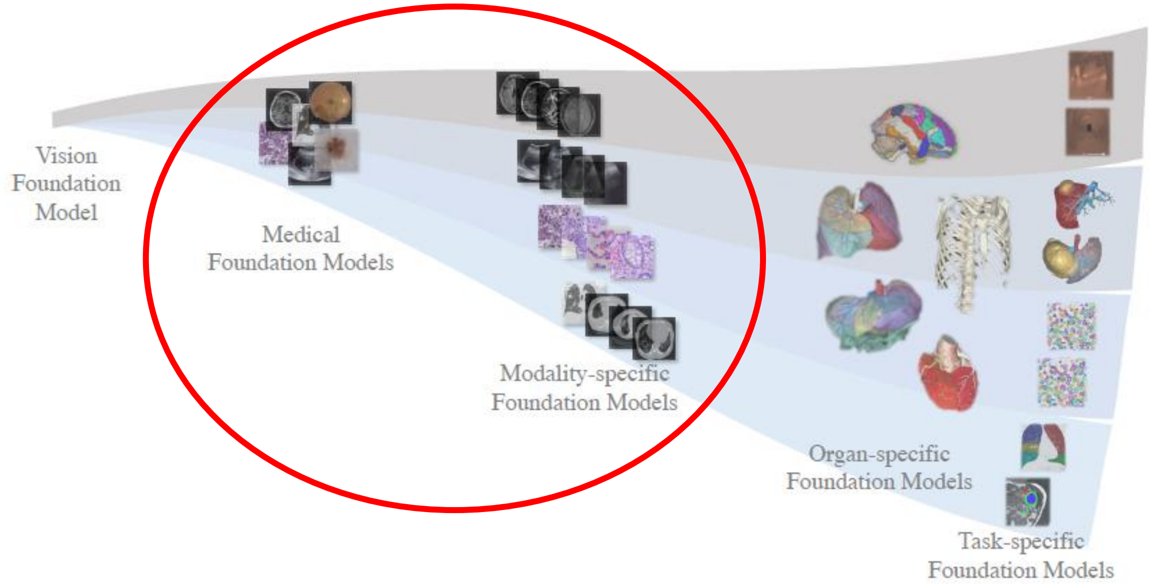
BUT… <u>On VLMs for classification it is the case.</u>

(a) *Zero-shot*		MESSIDOR	FIVES	REFUGE	20x3	$ODIR_{200x3}$	MMAC	**Avg.**
CLIP	ViT-B/32	0.200	0.256	0.433	0.333	0.480	0.183	0.314
BiomedCLIP	ViT-B/16	0.207	0.415	0.624	0.617	0.583	0.274	0.453
FLAIR	RN50	**0.604**	**0.735**	**0.883**	**0.983**	**0.667**	**0.400**	**0.712**
(b) *Linear Probing*								
ImageNet	RN50	0.424	0.741	0.733	0.983	0.887	0.631	0.733
CLIP	ViT-B/32	0.491	0.800	0.720	0.950	0.917	0.642	0.753
BiomedCLIP	ViT-B/16	0.433	0.654	0.776	0.866	0.883	0.678	0.715
RETFound	ViT-B/16	0.457	0.765	0.747	0.950	0.887	0.547	0.725
FLAIR	RN50	**0.719**	**0.879**	**0.843**	**1.000**	**0.935**	**0.740**	**0.852**

Silva-Rodríguez et *al*. A Foundation Language-Image Model of the Retina: Encoding Expert Knowledge in Text Supervision. MedIA’24.

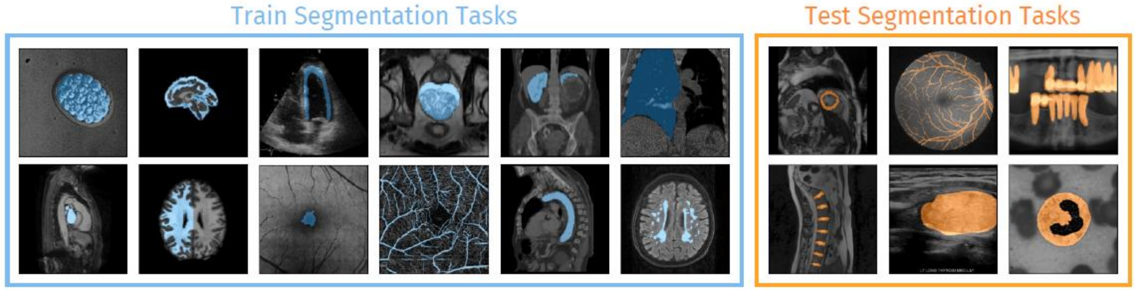
Types of foundation models: a data perspective.

Generalist *vs.* Specialized (pre-training)

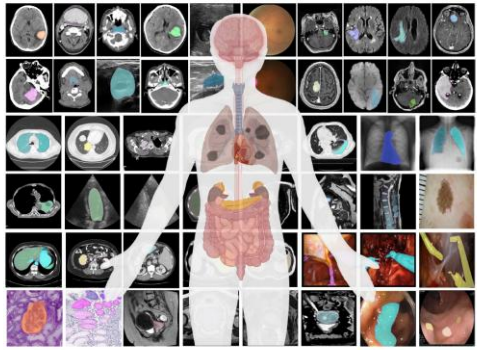


Huang et *al*. On The Challenges And Perspectives of Foundation Models For Medical Image Analysis. MedIA'24.

→ Modality better than Medical ?
(scarce empirical studies for segmentation)
BUT… <u>Large domain GAP between modalities</u>.



Butoi et al. Universeg: Universal medical image segmentation. ICCV'23.

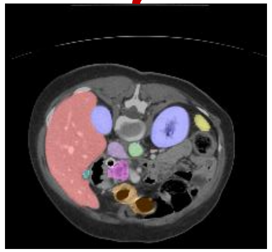


Ma et al. Segment Anything in Medical Images. Nat.Com.'24

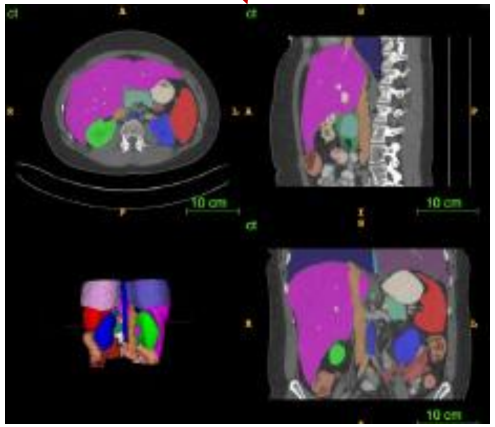
Types of foundation models: a data perspective.

2D *vs.* 3D (pre-training)

Actually…



2D Images*
256 x 256 pixels
512 x 512 pixels



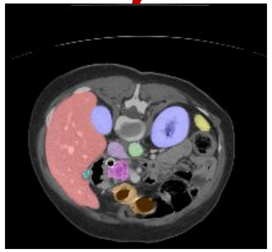
3D Volumes
256 x 256 x 500 pixels
512 x 512 x 500 pixels

* These scales not apply to other categories such as histology WSIs

Types of foundation models: a data perspective.

2D *vs.* 3D (pre-training)

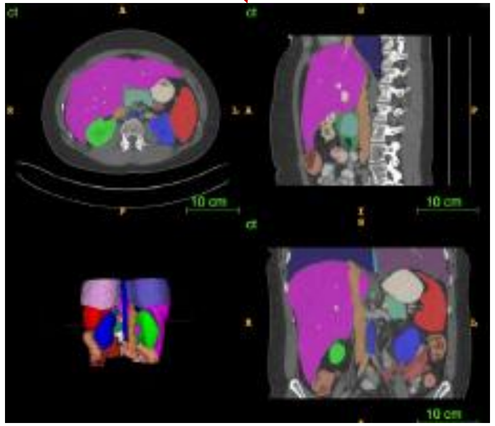
Actually…



2D Images*

256 x 256 pixels

512 x 512 pixels



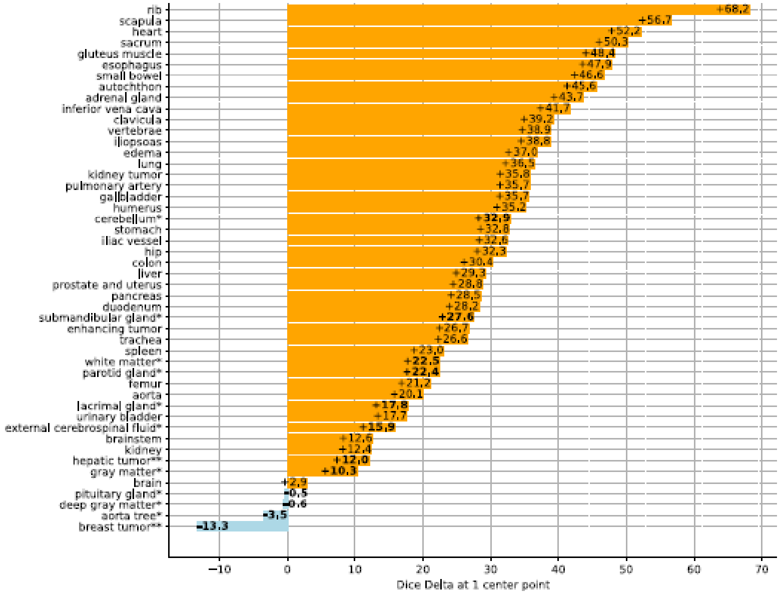
3D Volumes

256 x 256 x 500 pixels

512 x 512 x 500 pixels

→ Pre-training on 3D better than on 2D

(also, a limitation of natural image pre-training)



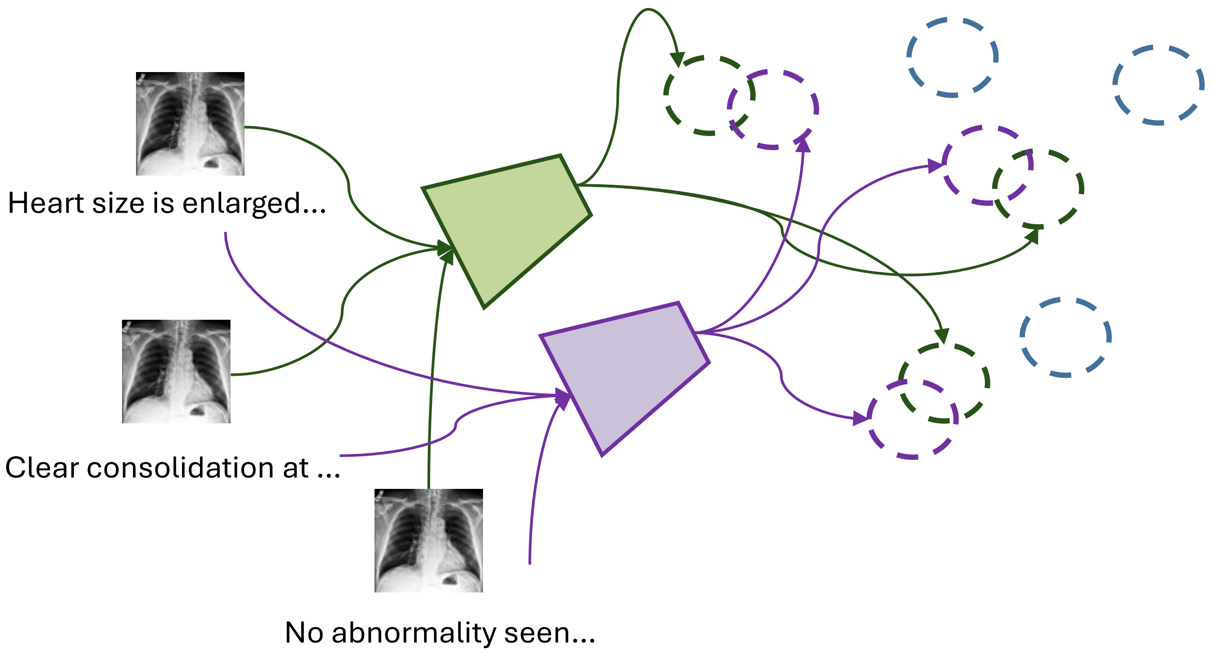
(d) SAM-Med3D *vs.* SAM-Med2D

Wang et al. SAM-Med3D: Towards General-Purpose Segmentation Models for Volumetric Medical Images. ArXiv'24.

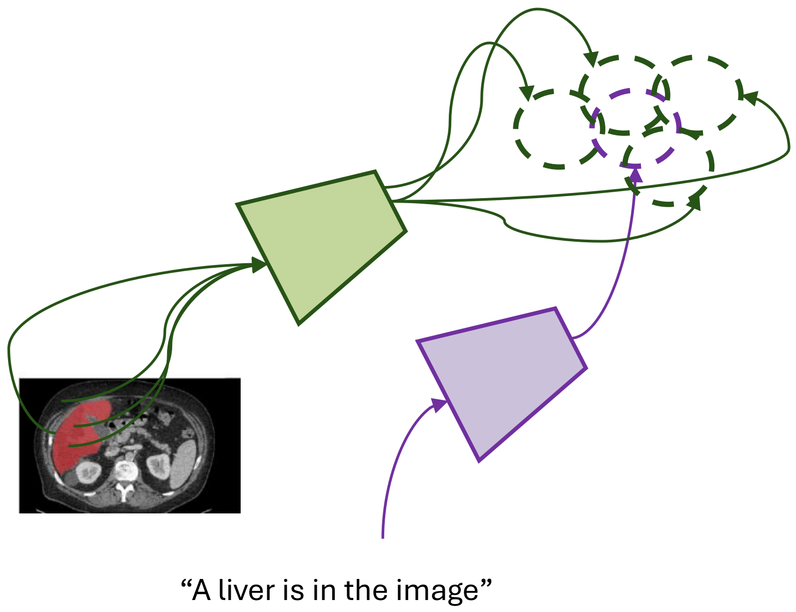
Types of foundation models: a data perspective.

Multimodal vs. Unimodal

Image-Level image-language pre-training



Segmentation image-language pre-training



Types of foundation models: a data perspective.

Multimodal vs. Unimodal

→ **Medical Image Segmentation Foundation Models are (so far) Unimodal** **(FMs are not necessary multi-modal)**

1. **Scarcity of grounding language annotations with masks.**

2. Already-existing large datasets with pixel/voxel annotations only.

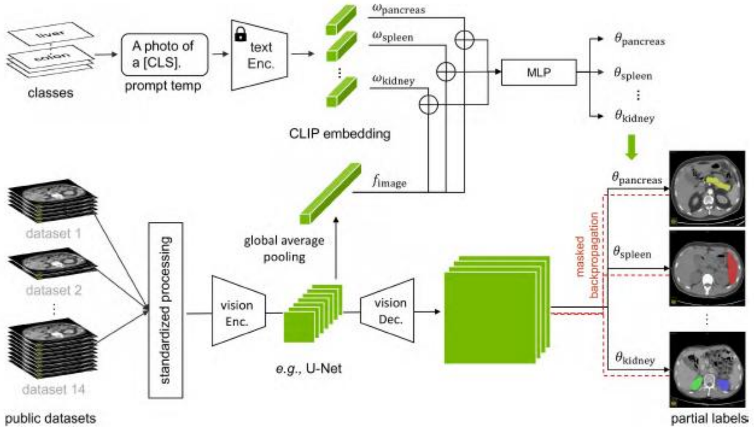
3. Unclear contribution of text modality in absence of open-vocabulary concepts.

"A liver is in the image"

4. Some works include a CLIP-driven component, but its contribution is doubtful. (We will see this latter)

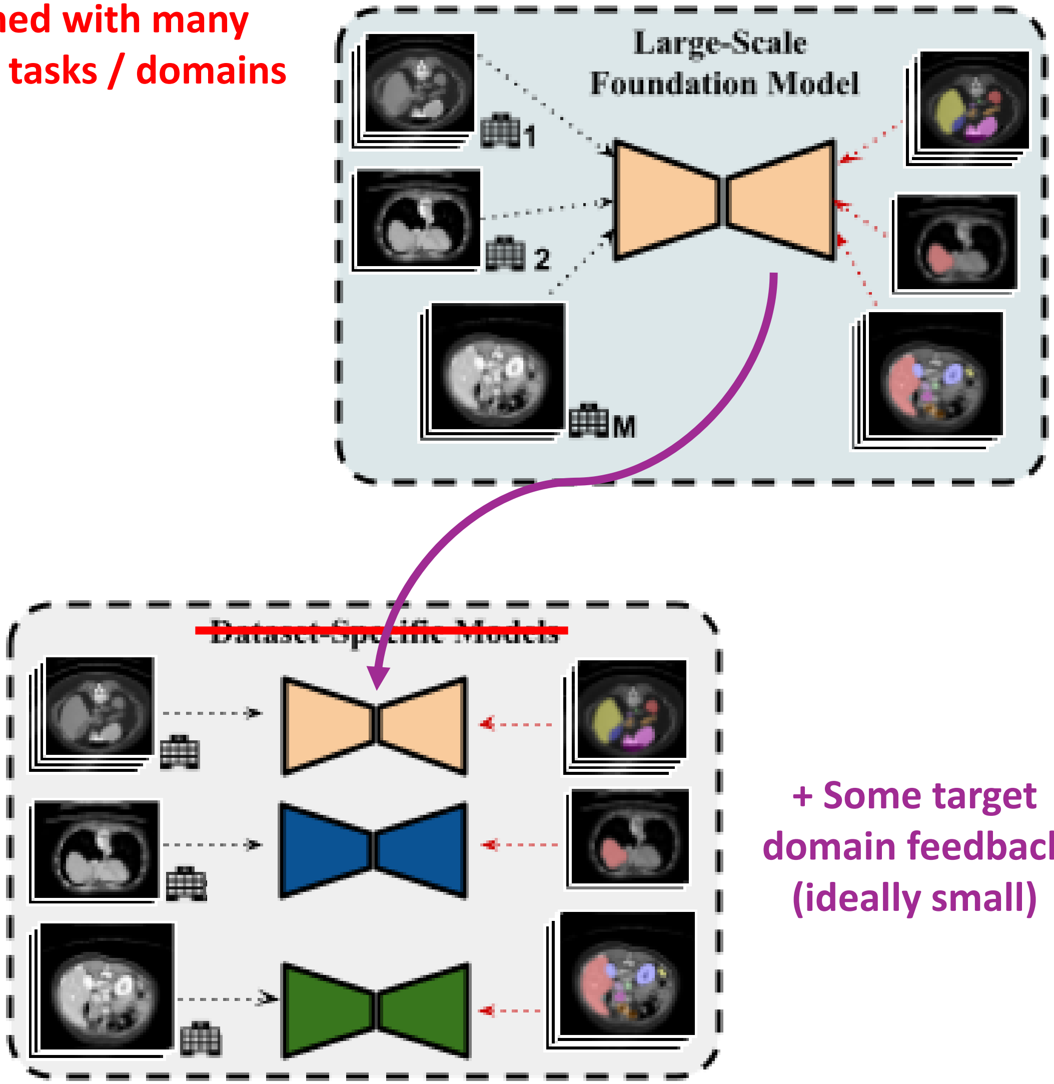
5. To explore in lesion segmentation?

Liu et al. CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection. ICCV'23.



Foundation models for medical image segmentation

Trained with many data / tasks / domains



+ Some target domain feedback (ideally small)

Organizing the mess!

1. Types of foundation models: a data perspective.
   - A. Generalist vs. Specialized
   - B. 2D *vs*. 3D
   - C. Multimodal *vs*. Unimodal
2. **Learning/Usage Objectives**
   - **A. Zero-shot / Transfer Learning**
   - **B. In-Context Learning**
   - **C. Interactive Models ("SAM")**
3. Zero-shot / Adaptation-oriented (3D data)
   - A. How to pre-train?
   - B. How useful are foundation models? Limitations on the adaptation stage
   - C. Few-shot Parameter-Efficient Fine-Tuning

Learning / usage objectives.

Med3D('19)

CLIP-Driven

MultiTalent

HERMES

UniSeg

FSEFT

SuPreM

Zero-shot / Transfer Learning

ImageNet Philosophy

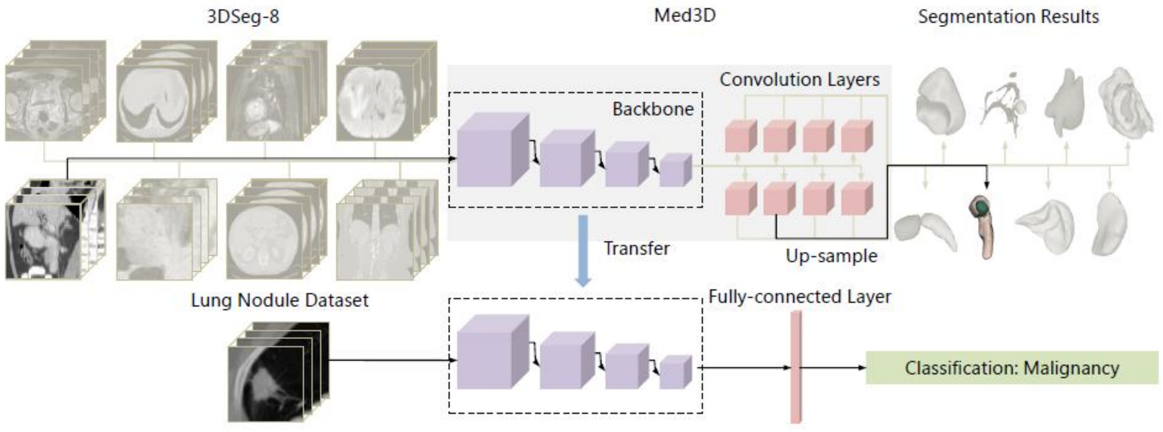
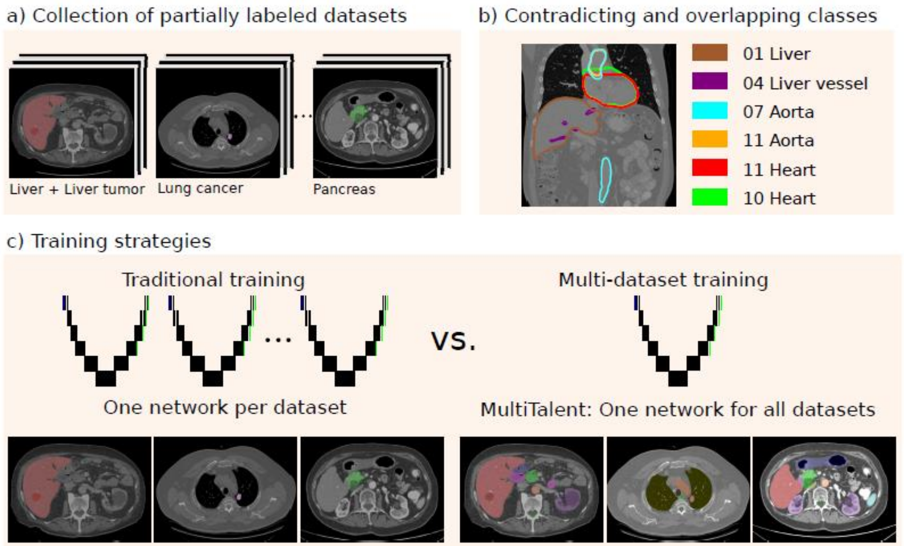


Figure 2: Framework of the proposed method.

Chen et *al*. Med3D: Transfer Learning for 3D Medical Image Analysis. ArXiv'19.



Ulrich et *al*. MultiTalent: A Multi-Dataset Approach to Medical Image Segmentation. MICCAI'23.

Learning / usage objectives.

Med3D('19)

CLIP-Driven

MultiTalent

HERMES

UniSeg

FSEFT

SuPreM

Zero-shot / Transfer Learning

ImageNet Philosophy

Zero-shot predictions to *base tasks*

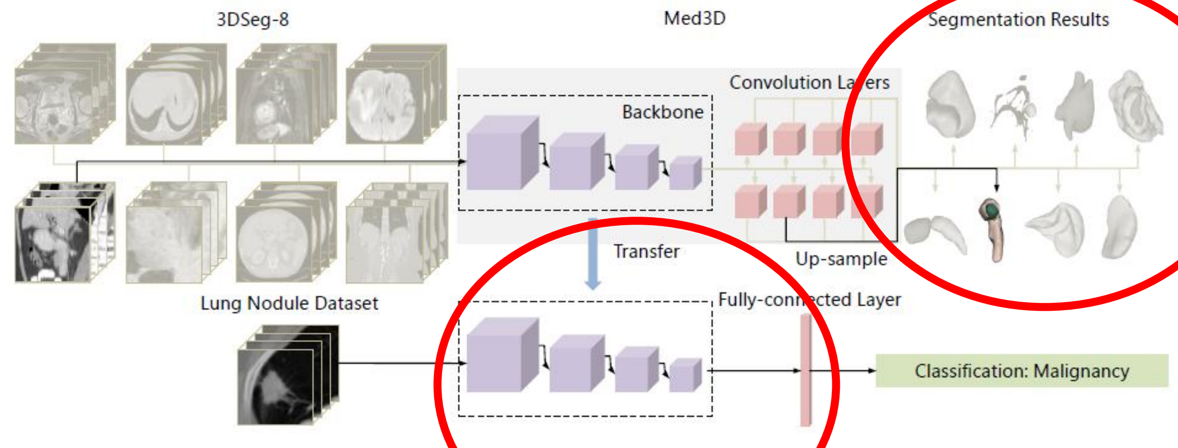
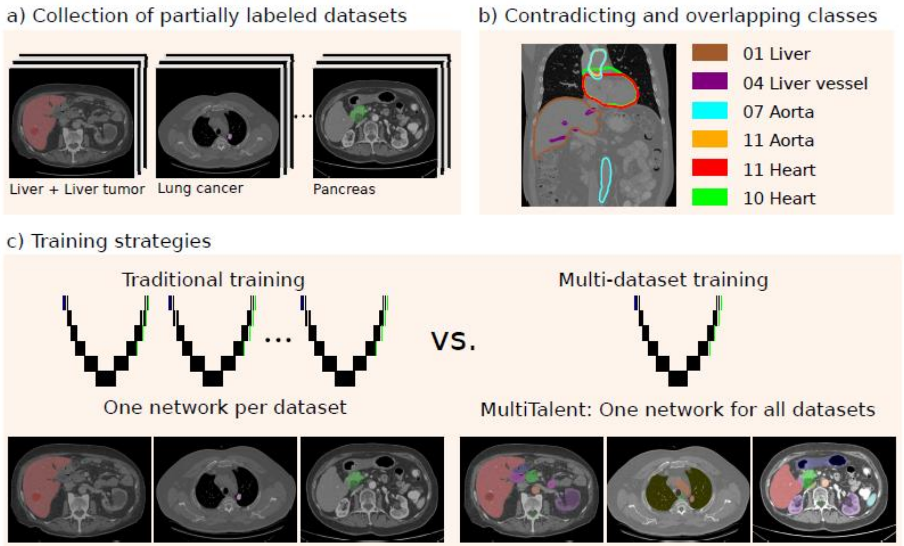


Figure 2: Framework of the proposed method.

Chen et *al*. Med3D: Transfer Learning for 3D Medical Image Analysis. ArXiv'19.

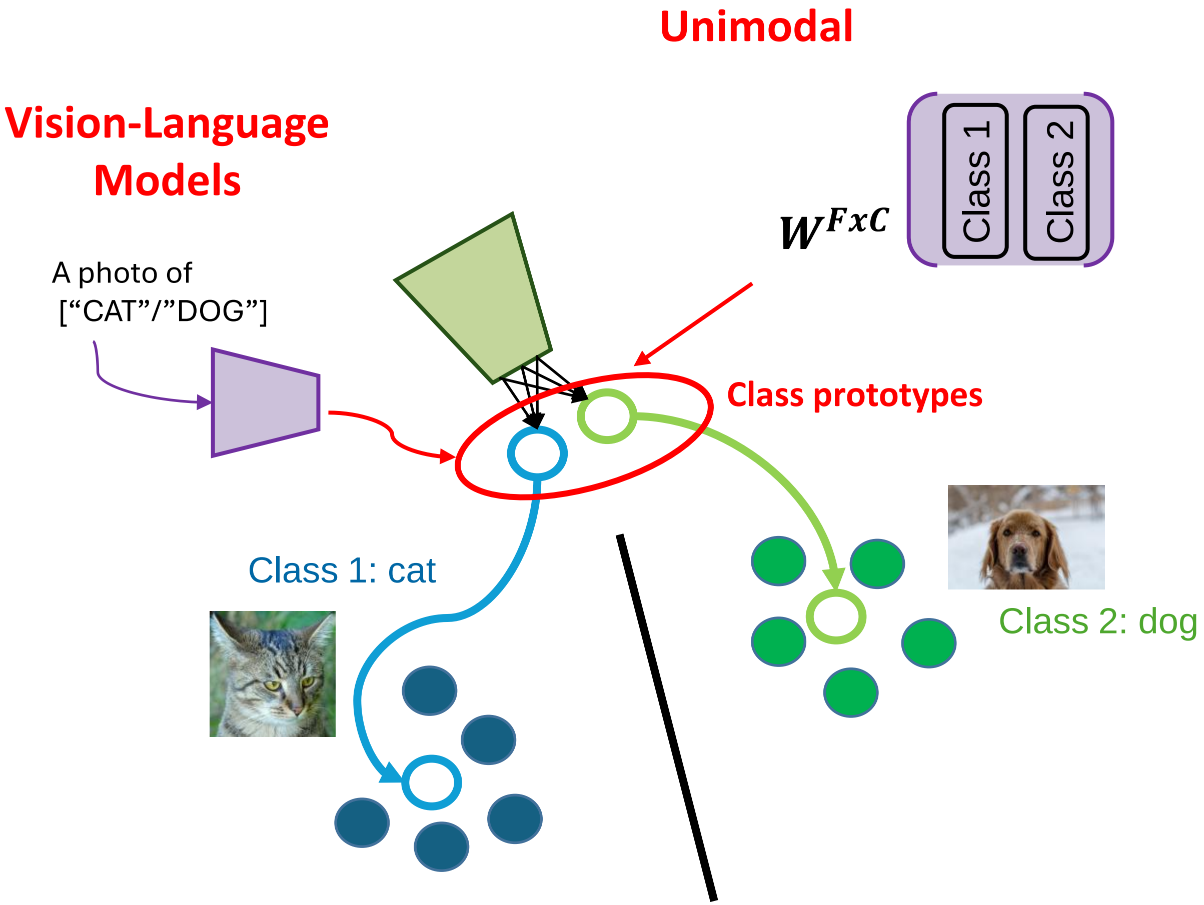
Fine-tuning to *novel domains/tasks*



Ulrich et *al*. MultiTalent: A Multi-Dataset Approach to Medical Image Segmentation. MICCAI'23.

Learning / usage objectives.

(Zero-shot: VLMs *vs.* Unimodal)



Zero-shot: not receiving any supervision from the target domain/task

Is zero-shot predictions to novel categories a realistic objective?

Undandarao et al. No Zero-Shot without Exponential Data: Pretraining Concept frequency Determines Multimodal Model Performance. ICLRW-FM'24.

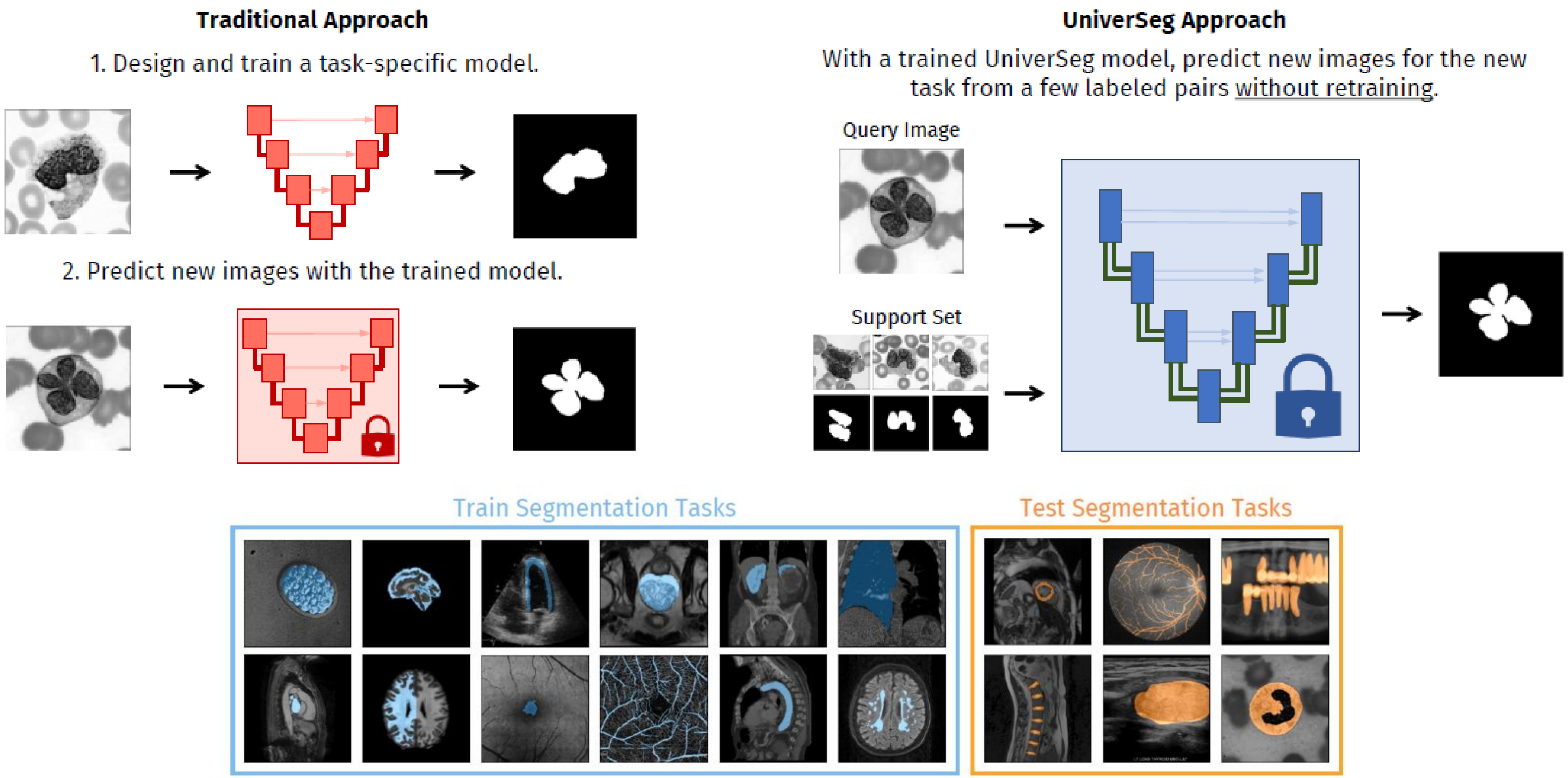
Learning / usage objectives.

UniverSeg

Tyche

In Context Learning

"At the end of the day, practitioners won't fine-tune"



Butoi et al. Universeg: Universal medical image segmentation. ICCV'23.

Learning / usage objectives.

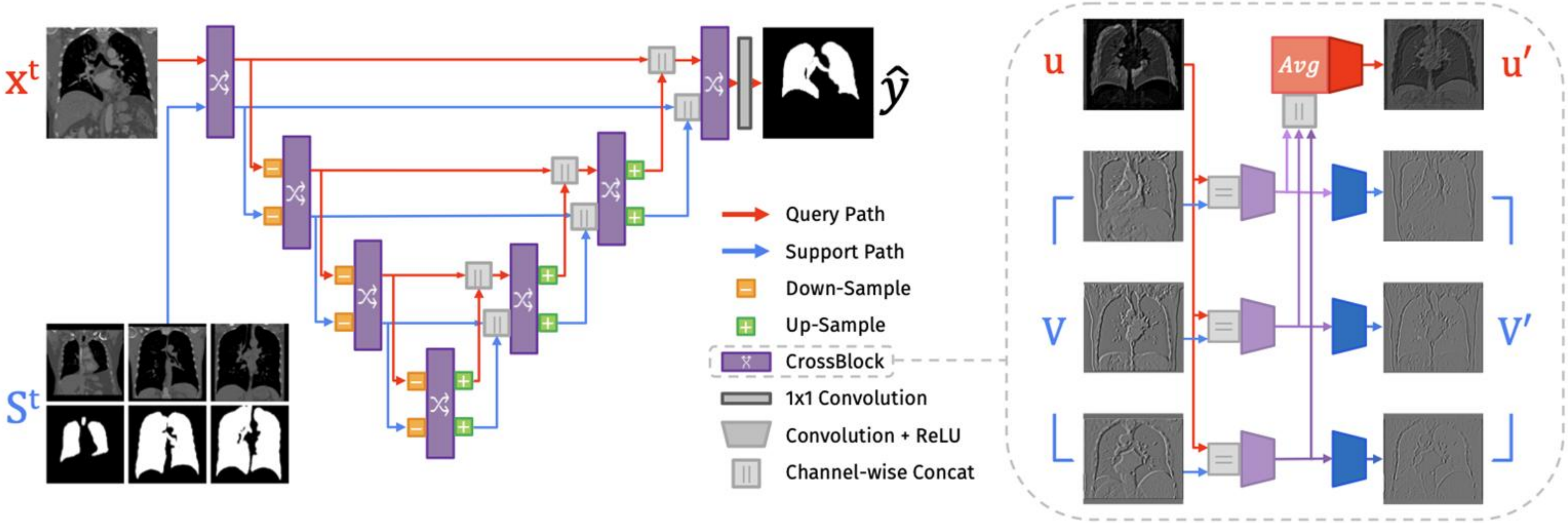
UniverSeg

In Context Learning

Main Idea

Query sample

Support set



Butoi et al. Universeg: Universal medical image segmentation. ICCV'23.

Learning / usage objectives.

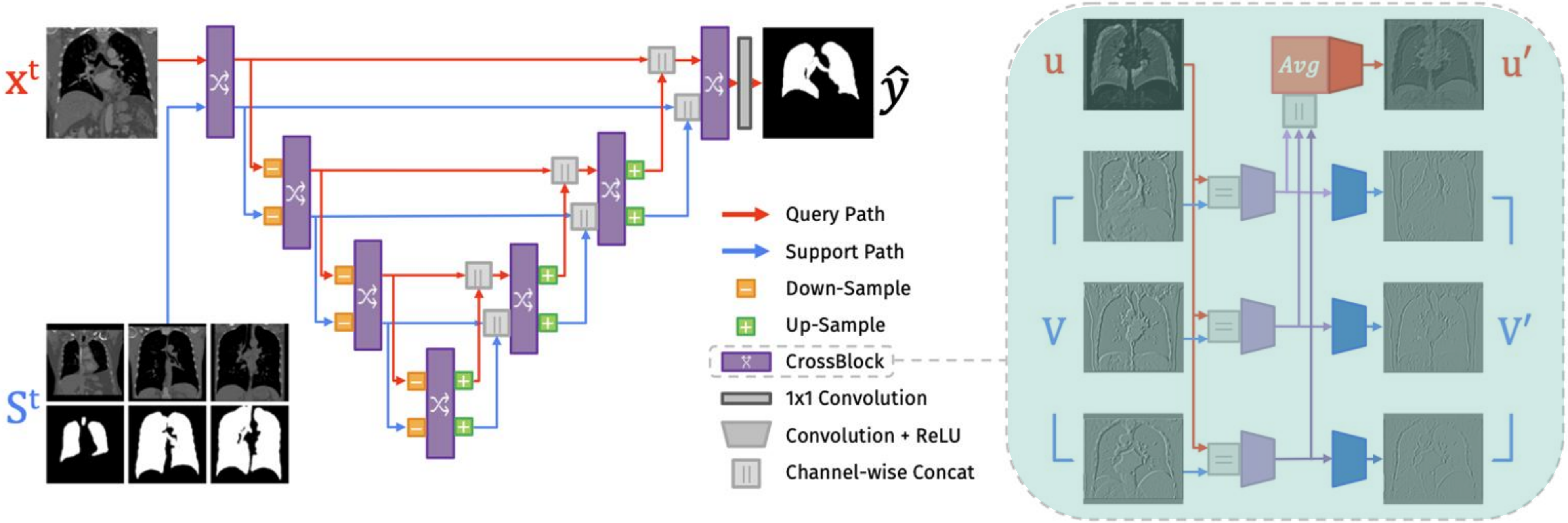
UniverSeg

In Context Learning

Main Idea

Query sample

Support set



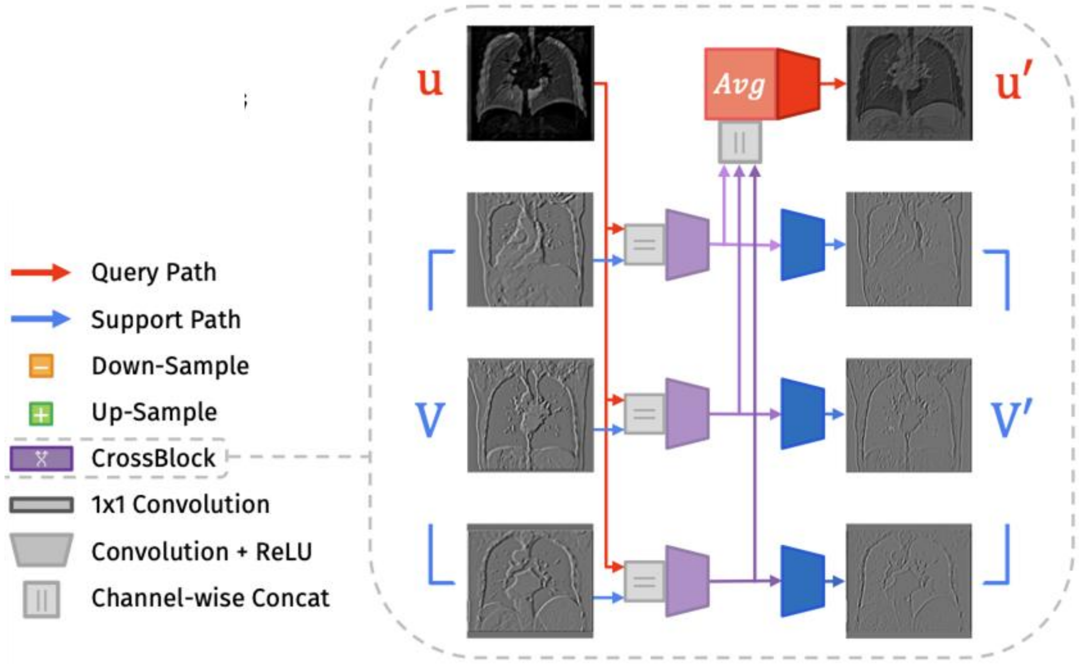
The representations from the query and support samples can interact at multiple scales

Butoi et al. Universeg: Universal medical image segmentation. ICCV'23.

Learning / usage objectives.

UniverSeg

In Context Learning



$$\mathrm{CrossBlock}(u, V; \theta_z, \theta_v) = (u', V'), \text{where:} \qquad (2)$$

$$z_i = A(\mathrm{CrossConv}(u, v_i; \theta_z)) \quad \text{for } i = 1, 2, \ldots, n$$

$$u' = 1/n \sum_{i=1}^{n} z_i$$

Query output: average across support

$$v'_i = A(\mathrm{Conv}(z_i; \theta_v)) \quad \text{for } i = 1, 2, \ldots, n,$$

Support samples activation maps

$$\mathrm{CrossConv}(u, V; \theta_z) = \{z_i\}_{i=1}^{n},$$
$$\text{for } z_i = \mathrm{Conv}(u || v_i; \theta_z),$$

Concatenate query and support activation maps

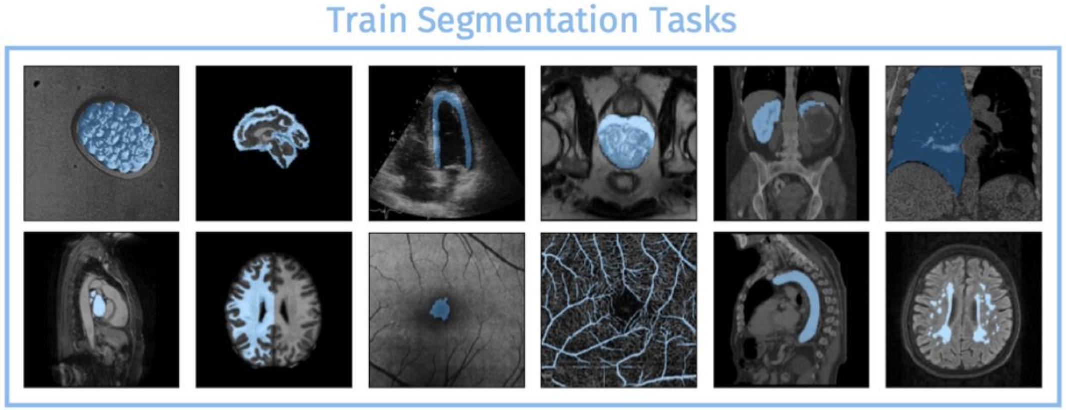
Butoi et al. Universeg: Universal medical image segmentation. ICCV'23.

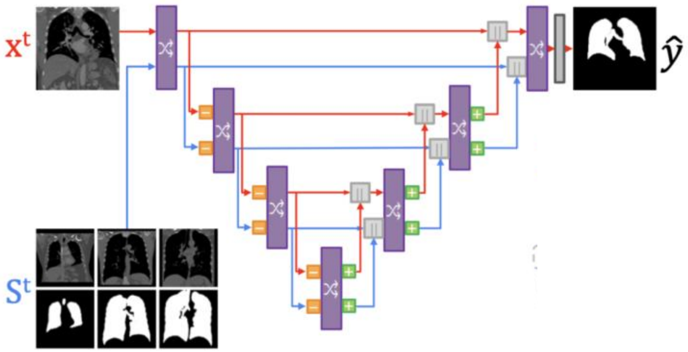
Learning / usage objectives.

UniverSeg

In Context Learning

How this is trained? (Hint: based on meta-learning or *learning-to-learn*)





Butoi et al. Universeg: Universal medical image segmentation. ICCV'23.

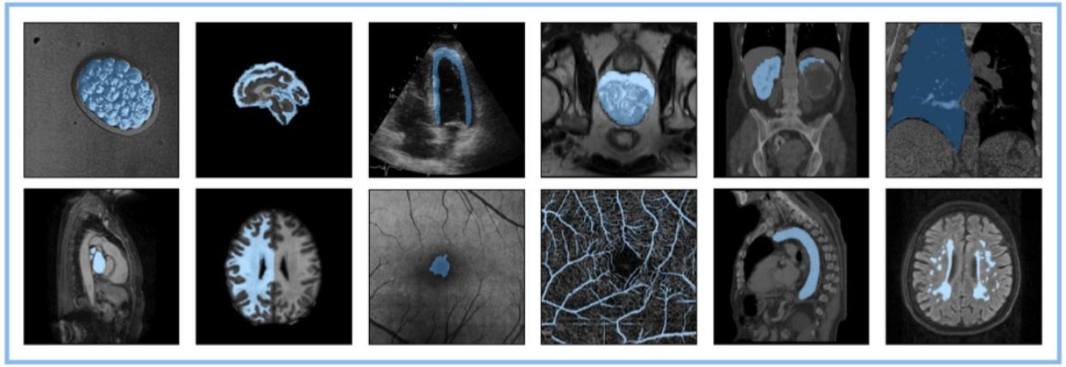
Learning / usage objectives.

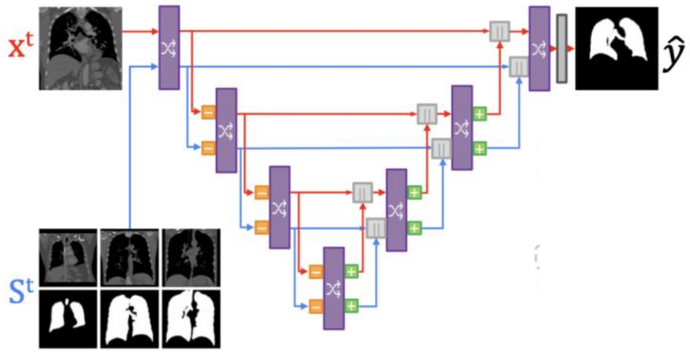
UniverSeg

In Context Learning

How this is trained? (Hint: based on meta-learning or *learning-to-learn*)

Train Segmentation Tasks





for $k = 1, \ldots, \text{NumTrainSteps}$ **do**

- $t \sim \mathcal{T}$ ▷ Sample Task
- $(x_i^t, y_i^t) \sim t$ ▷ Sample Query
- $S^t \leftarrow \{(x_j^t, y_j^t)\}_{j \neq i}^n$ ▷ Sample Support
- $x_i^t, y_i^t \leftarrow \text{Aug}_t(x_i^t, y_i^t)$ ▷ Augment Query
- $S^t \leftarrow \{\text{Aug}_t(x_j^t, y_j^t)\}_j^n$ ▷ Augment Support
- $x_i^t, y_i^t, S^t \leftarrow \text{Aug}_T(x_i^t, y_i^t, S^t)$ ▷ Task Aug
- $\hat{y}_i \leftarrow f_\theta(x_i^t, S^t)$ ▷ Predict label map
- $\ell \leftarrow \mathcal{L}_{\text{seg}}(\hat{y}_i, y_i^t)$ ▷ Compute loss
- $\theta \leftarrow \theta - \eta \nabla_\theta \ell$ ▷ Gradient step

end for

Butoi et al. Universeg: Universal medical image segmentation. ICCV'23.

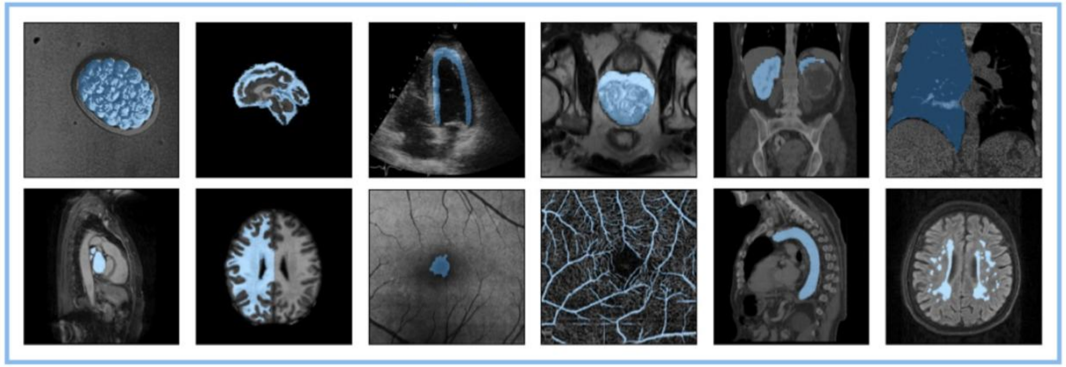
Learning / usage objectives.

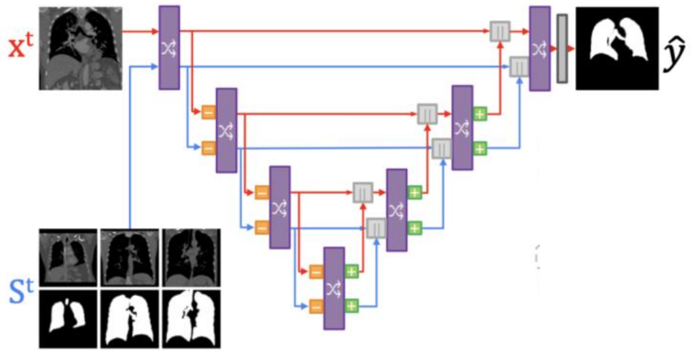
UniverSeg

In Context Learning

How this is trained? (Hint: based on meta-learning or *learning-to-learn*)

Train Segmentation Tasks





for $k = 1, \ldots, \text{NumTrainSteps}$ **do**

- $t \sim \mathcal{T}$ ▷ Sample Task
- $(x_i^t, y_i^t) \sim t$ ▷ Sample Query
- $S^t \leftarrow \{(x_j^t, y_j^t)\}_{j \neq i}^n$ ▷ Sample Support
- $x_i^t, y_i^t \leftarrow \text{Aug}_t(x_i^t, y_i^t)$ ▷ Augment Query
- $S^t \leftarrow \{\text{Aug}_t(x_j^t, y_j^t)\}_j^n$ ▷ Augment Support
- $x_i^t, y_i^t, S^t \leftarrow \text{Aug}_T(x_i^t, y_i^t, S^t)$ ▷ Task Aug
- $\hat{y}_i \leftarrow f_\theta(x_i^t, S^t)$ ▷ Predict label map
- $\ell \leftarrow \mathcal{L}_{\text{seg}}(\hat{y}_i, y_i^t)$ ▷ Compute loss
- $\theta \leftarrow \theta - \eta \nabla_\theta \ell$ ▷ Gradient step

end for

Among all training tasks

Butoi et al. Universeg: Universal medical image segmentation. ICCV'23.

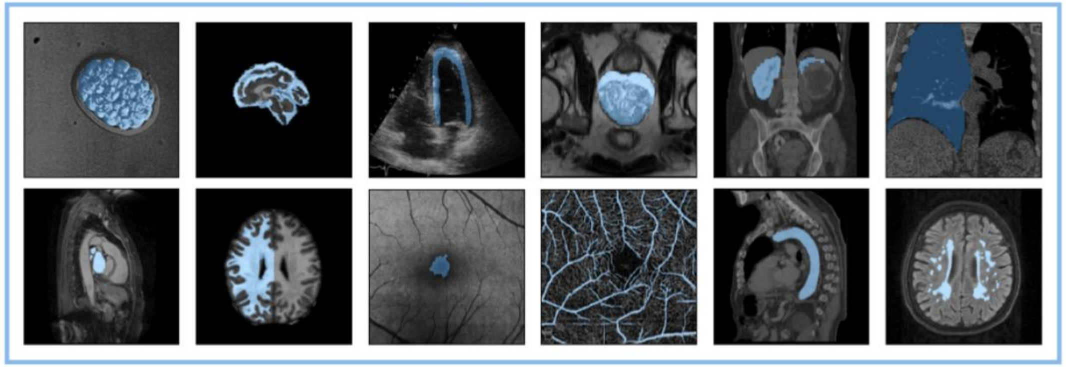
Learning / usage objectives.

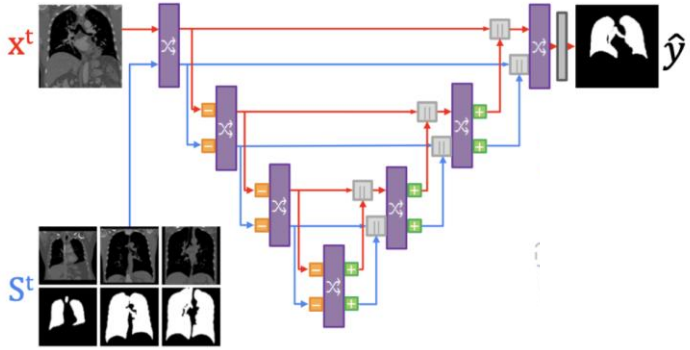
UniverSeg

In Context Learning

How this is trained? (Hint: based on meta-learning or *learning-to-learn*)

Train Segmentation Tasks



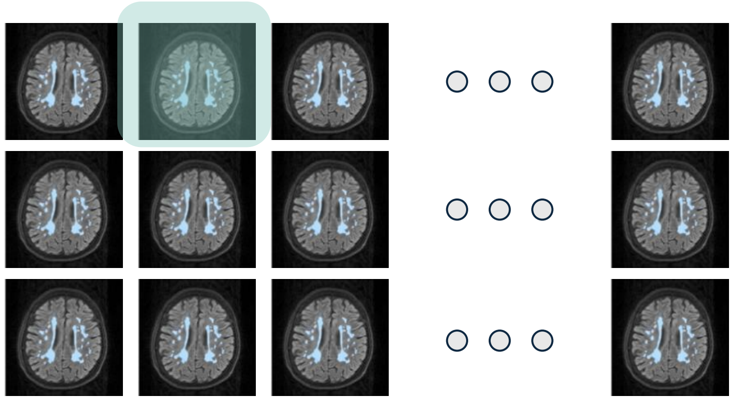


for $k = 1, \ldots, \text{NumTrainSteps}$ **do**

- $t \sim \mathcal{T}$ ▷ Sample Task
- $(x_i^t, y_i^t) \sim t$ ▷ Sample Query
- $S^t \leftarrow \{(x_j^t, y_j^t)\}_{j \neq i}^n$ ▷ Sample Support
- $x_i^t, y_i^t \leftarrow \text{Aug}_t(x_i^t, y_i^t)$ ▷ Augment Query
- $S^t \leftarrow \{\text{Aug}_t(x_j^t, y_j^t)\}_j^n$ ▷ Augment Support
- $x_i^t, y_i^t, S^t \leftarrow \text{Aug}_T(x_i^t, y_i^t, S^t)$ ▷ Task Aug
- $\hat{y}_i \leftarrow f_\theta(x_i^t, S^t)$ ▷ Predict label map
- $\ell \leftarrow \mathcal{L}_{\text{seg}}(\hat{y}_i, y_i^t)$ ▷ Compute loss
- $\theta \leftarrow \theta - \eta \nabla_\theta \ell$ ▷ Gradient step

end for

Among all training samples from that task



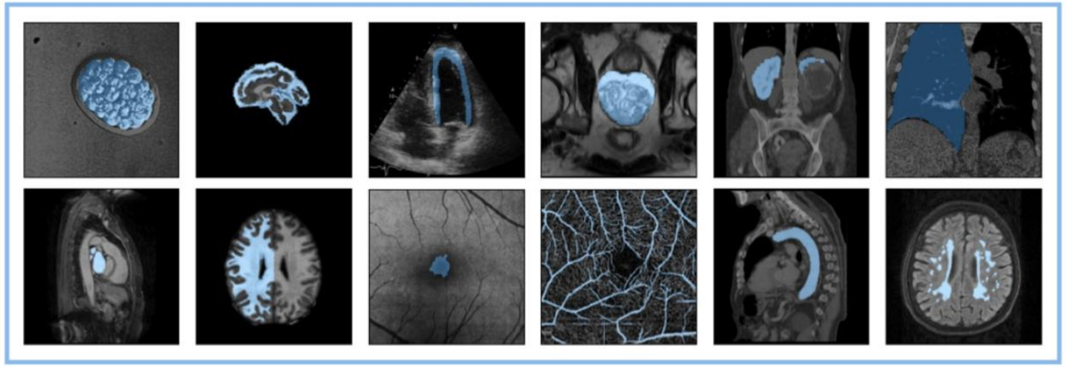
Butoi et al. Universeg: Universal medical image segmentation. ICCV'23.

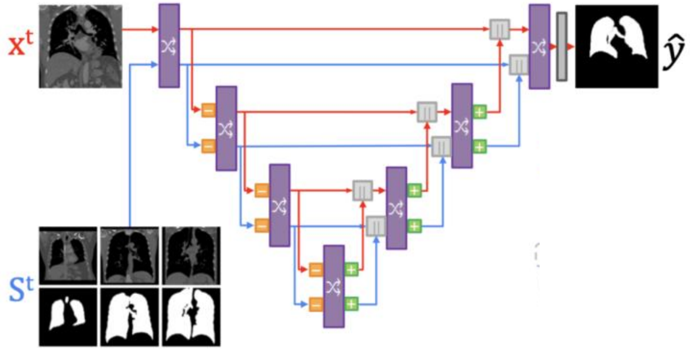
Learning / usage objectives.

In Context Learning

How this is trained? (Hint: based on meta-learning or *learning-to-learn*)

Train Segmentation Tasks



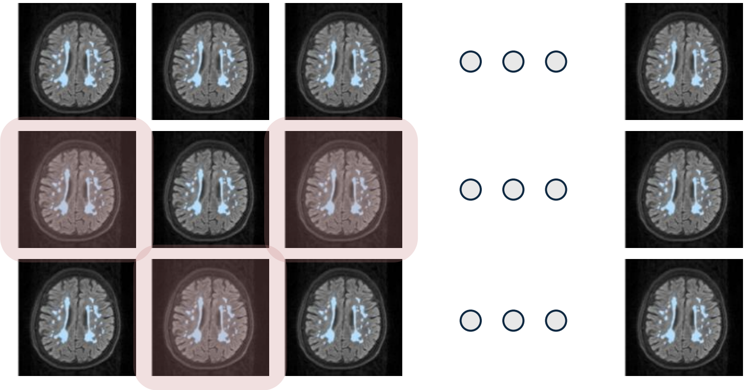


for $k = 1, \ldots, \text{NumTrainSteps}$ **do**

- $t \sim \mathcal{T}$ ▷ Sample Task
- $(x_i^t, y_i^t) \sim t$ ▷ Sample Query
- $S^t \leftarrow \{(x_j^t, y_j^t)\}_{j \neq i}^n$ ▷ Sample Support
- $x_i^t, y_i^t \leftarrow \text{Aug}_t(x_i^t, y_i^t)$ ▷ Augment Query
- $S^t \leftarrow \{\text{Aug}_t(x_j^t, y_j^t)\}_j^n$ ▷ Augment Support
- $x_i^t, y_i^t, S^t \leftarrow \text{Aug}_T(x_i^t, y_i^t, S^t)$ ▷ Task Aug
- $\hat{y}_i \leftarrow f_\theta(x_i^t, S^t)$ ▷ Predict label map
- $\ell \leftarrow \mathcal{L}_{\text{seg}}(\hat{y}_i, y_i^t)$ ▷ Compute loss
- $\theta \leftarrow \theta - \eta \nabla_\theta \ell$ ▷ Gradient step

end for

Among all training samples from that task



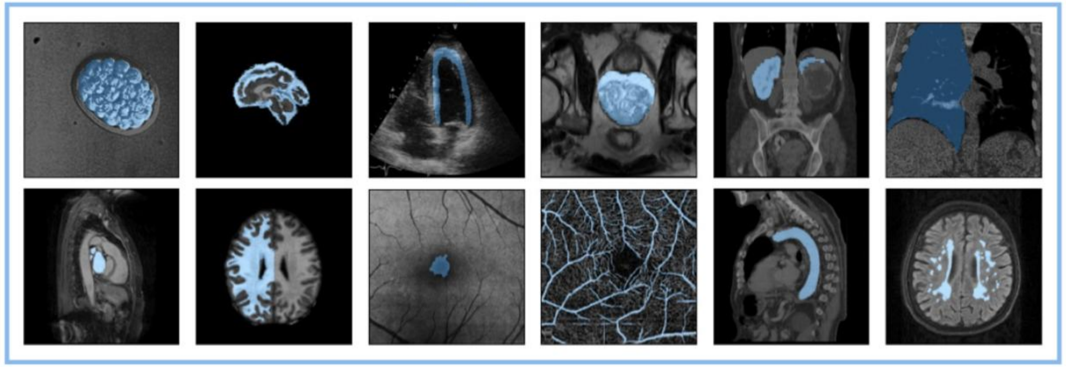
Butoi et al. Universeg: Universal medical image segmentation. ICCV'23.

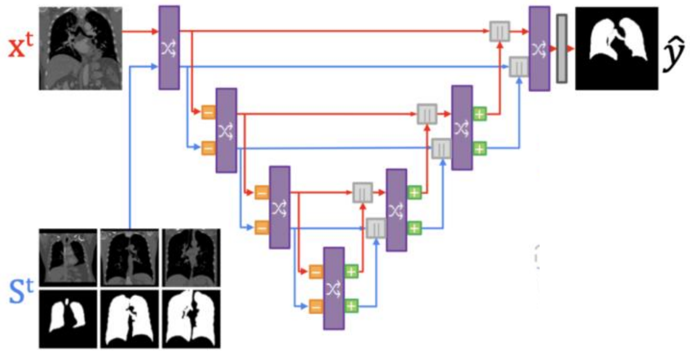
Learning / usage objectives.

In Context Learning

How this is trained? (Hint: based on meta-learning or *learning-to-learn*)

Train Segmentation Tasks





for $k = 1, \ldots, \text{NumTrainSteps}$ **do**

$t \sim \mathcal{T}$ ▷ Sample Task

$(x_i^t, y_i^t) \sim t$ ▷ Sample Query

$S^t \leftarrow \{(x_j^t, y_j^t)\}_{j \neq i}^n$ ▷ Sample Support

$x_i^t, y_i^t \leftarrow \text{Aug}_t(x_i^t, y_i^t)$ ▷ Augment Query

$S^t \leftarrow \{\text{Aug}_t(x_j^t, y_j^t)\}_j^n$ ▷ Augment Support

$x_i^t, y_i^t, S^t \leftarrow \text{Aug}_T(x_i^t, y_i^t, S^t)$ ▷ Task Aug

$\hat{y}_i \leftarrow f_\theta(x_i^t, S^t)$ ▷ Predict label map

$\ell \leftarrow \mathcal{L}_{\text{seg}}(\hat{y}_i, y_i^t)$ ▷ Compute loss

$\theta \leftarrow \theta - \eta \nabla_\theta \ell$ ▷ Gradient step

end for

→ Images Augmentations

Butoi et al. Universeg: Universal medical image segmentation. ICCV'23.

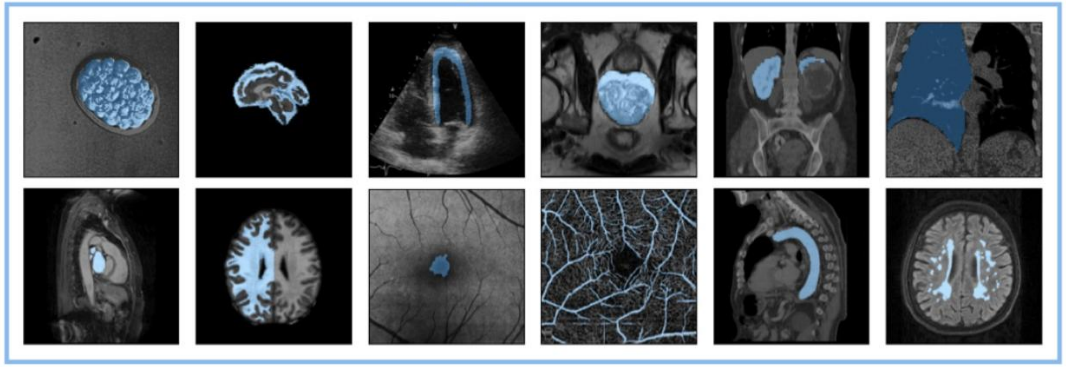
Learning / usage objectives.

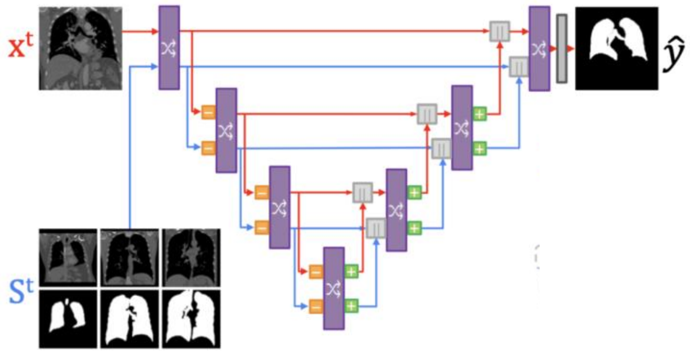
UniverSeg

In Context Learning

How this is trained? (Hint: based on meta-learning or *learning-to-learn*)

Train Segmentation Tasks





for $k = 1, \ldots, \text{NumTrainSteps}$ **do**

$t \sim \mathcal{T}$ ▷ Sample Task

$(x_i^t, y_i^t) \sim t$ ▷ Sample Query

$S^t \leftarrow \{(x_j^t, y_j^t)\}_{j \neq i}^n$ ▷ Sample Support

$x_i^t, y_i^t \leftarrow \text{Aug}_t(x_i^t, y_i^t)$ ▷ Augment Query

$S^t \leftarrow \{\text{Aug}_t(x_j^t, y_j^t)\}_j^n$ ▷ Augment Support

$x_i^t, y_i^t, S^t \leftarrow \text{Aug}_T(x_i^t, y_i^t, S^t)$ ▷ Task Aug

$\hat{y}_i \leftarrow f_\theta(x_i^t, S^t)$ ▷ Predict label map

$\ell \leftarrow \mathcal{L}_{\text{seg}}(\hat{y}_i, y_i^t)$ ▷ Compute loss

$\theta \leftarrow \theta - \eta \nabla_\theta \ell$ ▷ Gradient step

end for

Standard (training) forward-backward steps

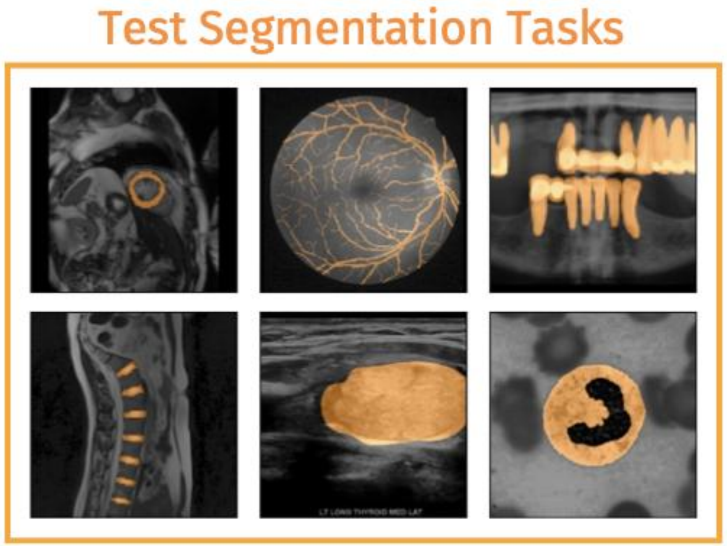
Butoi et al. Universeg: Universal medical image segmentation. ICCV'23.

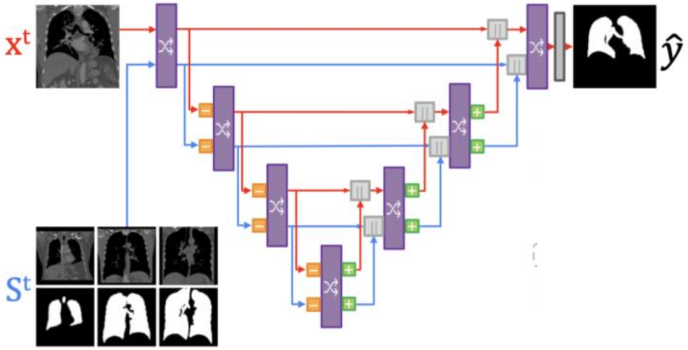
Learning / usage objectives.

UniverSeg

In Context Learning

And what about inference?





For a given image x^t $\hat{y} = f_\theta(x^t, S^t)$

To make it more robust, multiple support sets are employed

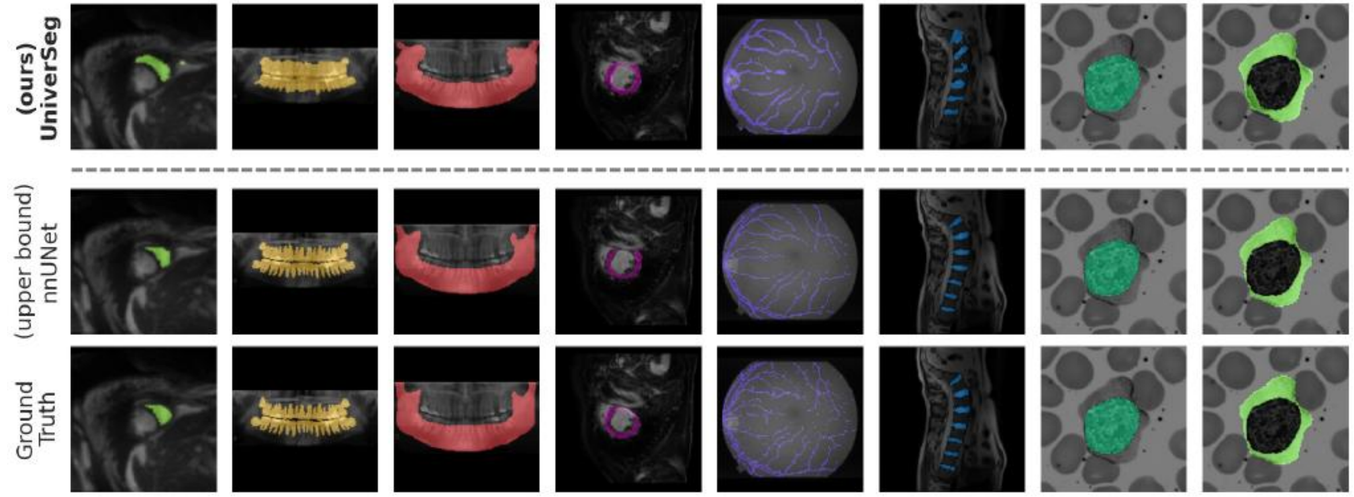
$$\hat{y} = \frac{1}{M}\sum_{m=1}^{M} f_\theta(x^t, S^t_m)$$

Butoi et al. Universeg: Universal medical image segmentation. ICCV'23.

Learning / usage objectives.

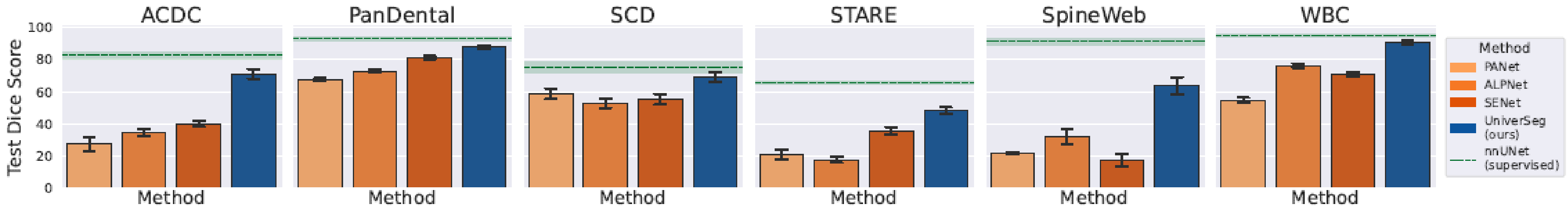
UniverSeg

In Context Learning



→ Can tackle new tasks.
→ Does not require fine-tuning.
→ Promising performance.

→ So far binary scenario.
→ Performance below dataset-specific models.
→ Unclear implementation on large 3D data.
→ Requires continuously employing the support set.



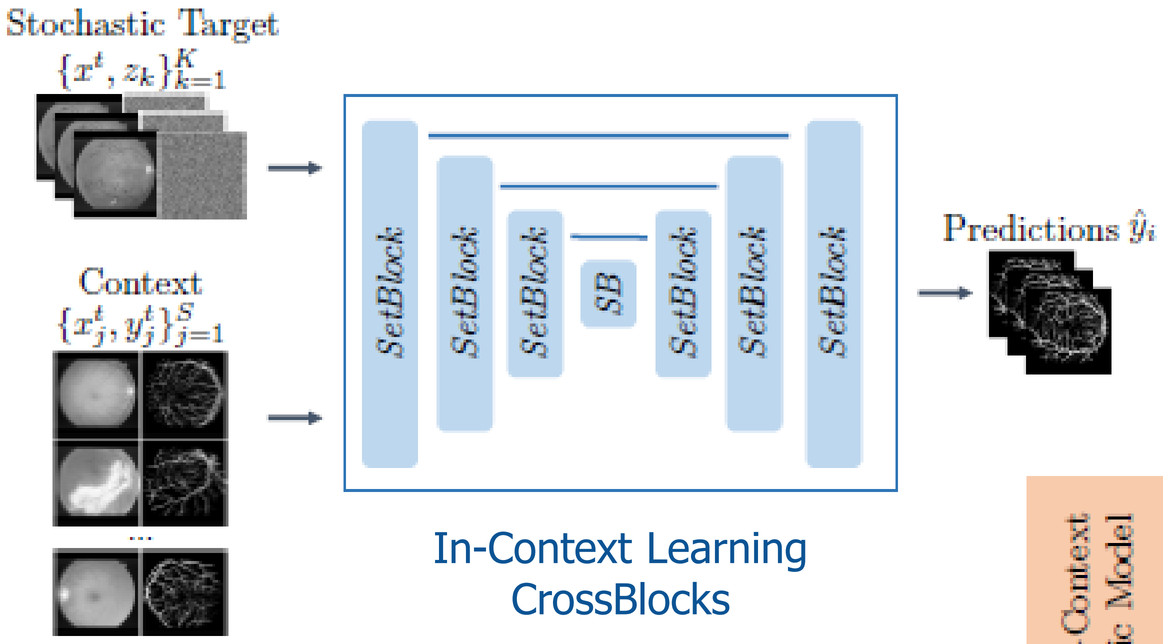
Butoi et al. Universeg: Universal medical image segmentation. ICCV'23.

Learning / usage objectives.

Tyche

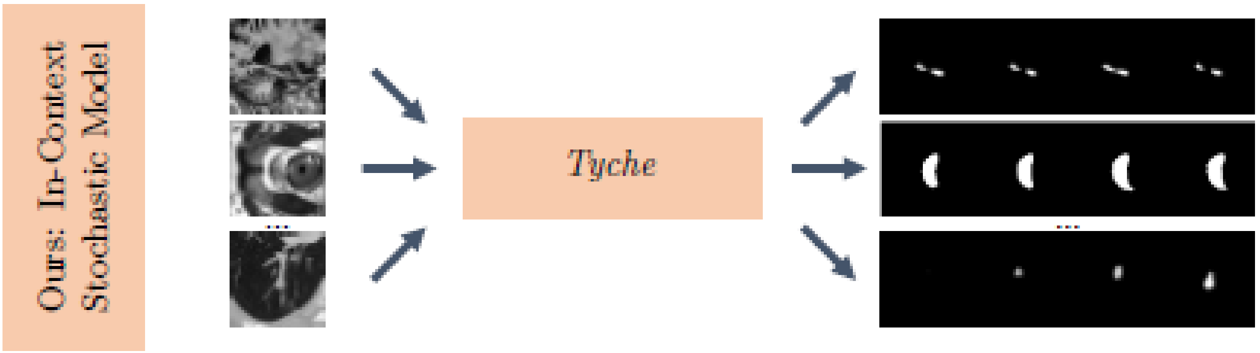
In Context Learning

Test-Time Augmentations



In-Context Learning CrossBlocks

Measure Uncertainty



Rakic et *al*. Tyche: Stochastic In-Context Learning for Medical Image Segmentation. CVPR'24.

Learning / usage objectives.

SAM

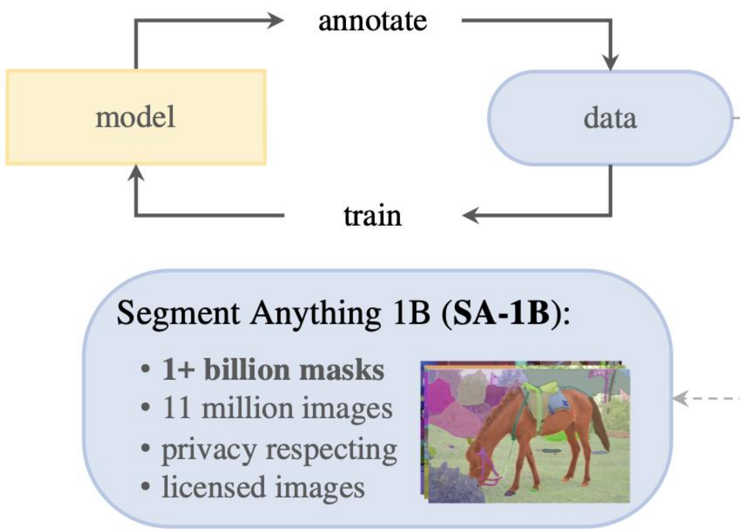
MedSAM

3DSAM-Adapter

MA-SAM

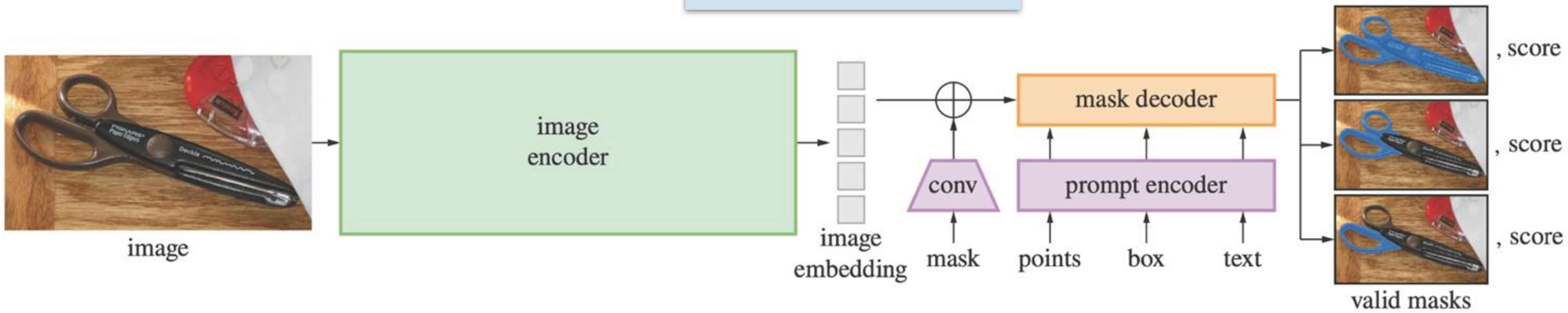
Med-SAM3D

Interactive models (“SAM”)



Trained on an huge amount of data

Pipeline



Kirillov et *al*. Segment Anything. ICCV’23.

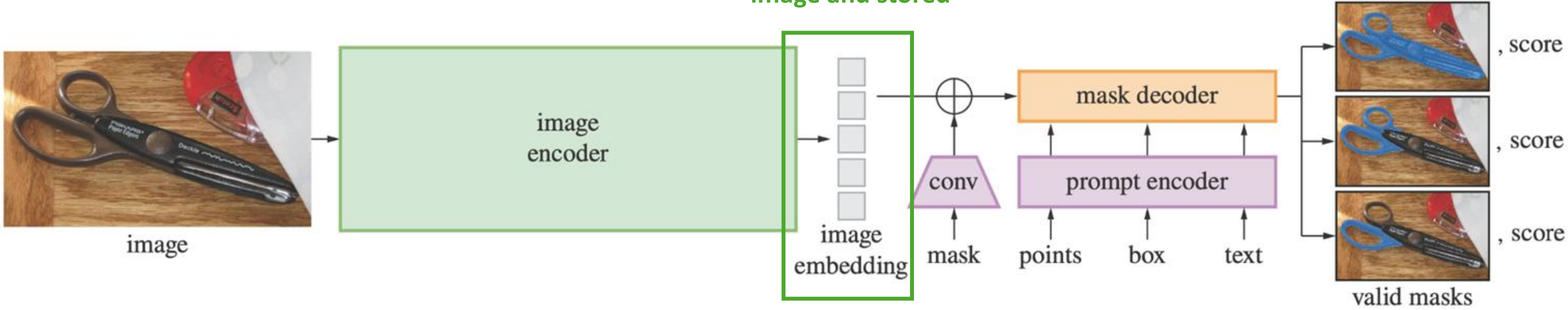
Learning / usage objectives.

SAM

Interactive models ("SAM")

How this is trained?

Computed once per image and stored



Kirillov et *al*. Segment Anything. ICCV'23.

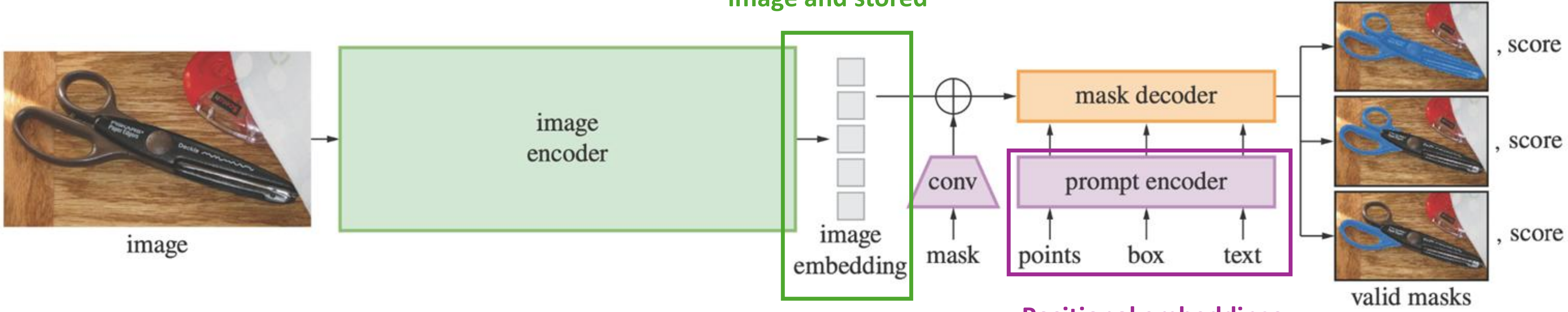
Learning / usage objectives.

SAM

Interactive models (“SAM”)

How this is trained?

Computed once per image and stored



Positional embeddings
+
Learnable Embeddings
(per type of prompt)

Kirillov et *al*. Segment Anything. ICCV’23.

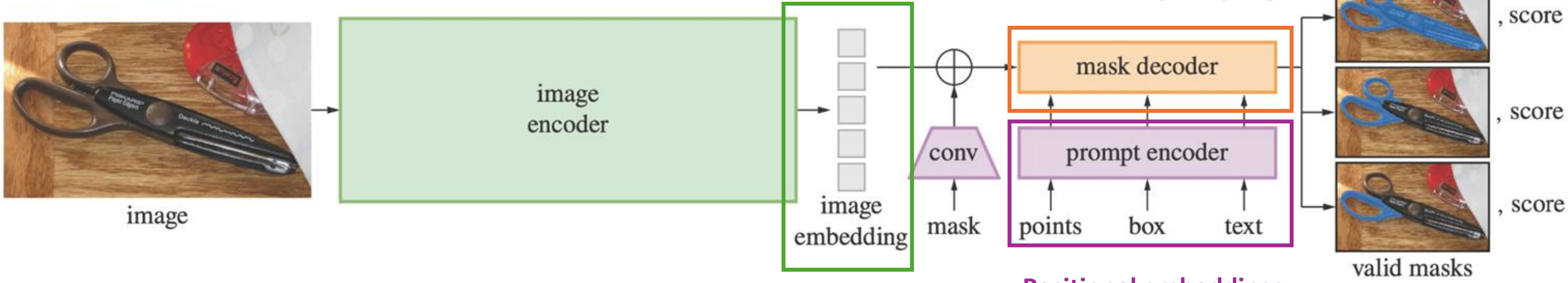
Learning / usage objectives.

Interactive models ("SAM")

How this is trained?

Computed once per image and stored

Lightweight Mask Decoder: Image – prompt cross-attention + MLP + upsampling



Positional embeddings
+
Learnable Embeddings
(per type of prompt)

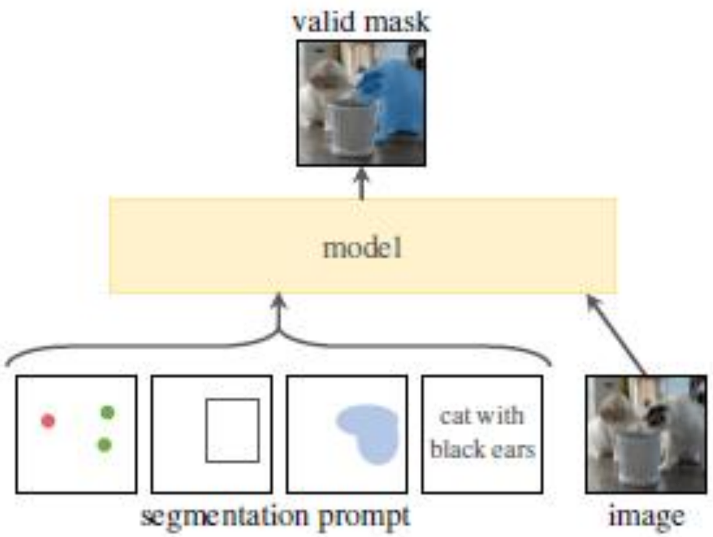
Kirillov et *al*. Segment Anything. ICCV'23.

Learning / usage objectives.

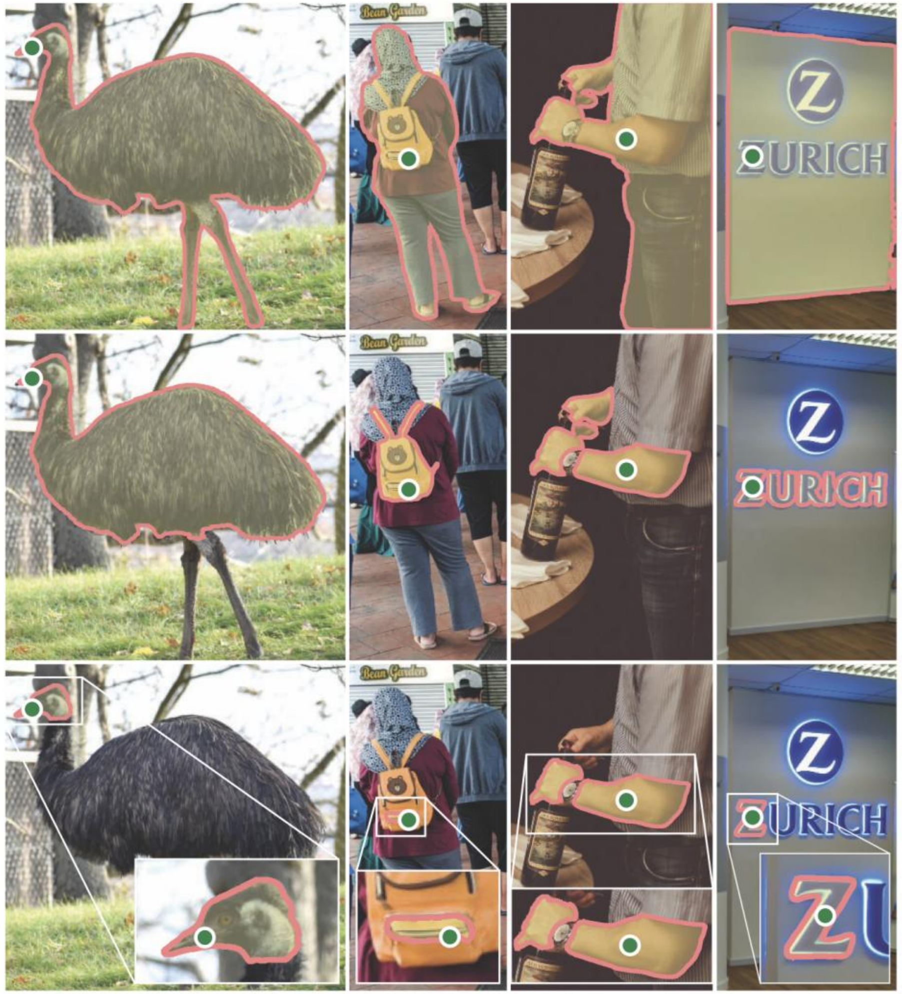
SAM

Interactive models ("SAM")

And what about inference?



Remember: prompts on test data

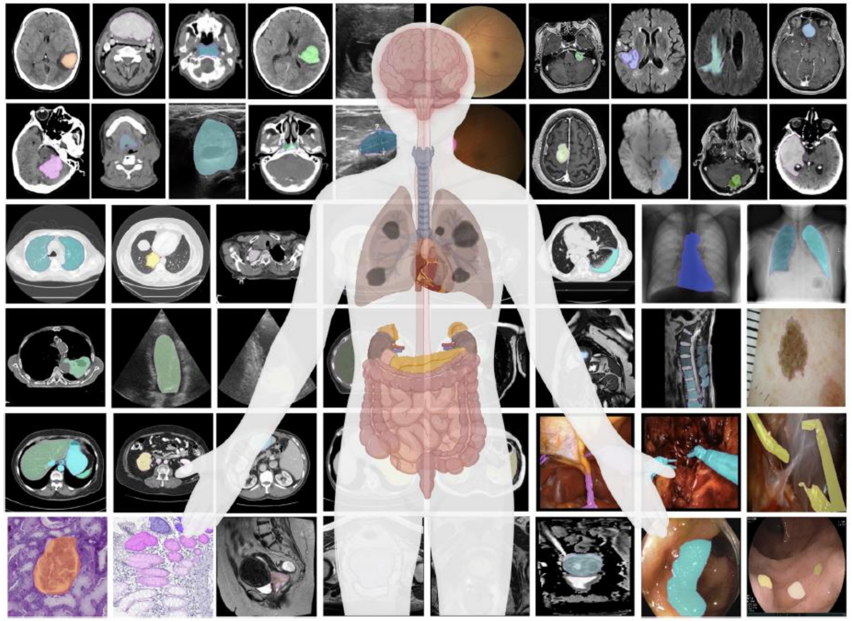


Kirillov et *al*. Segment Anything. ICCV'23.

Learning / usage objectives.

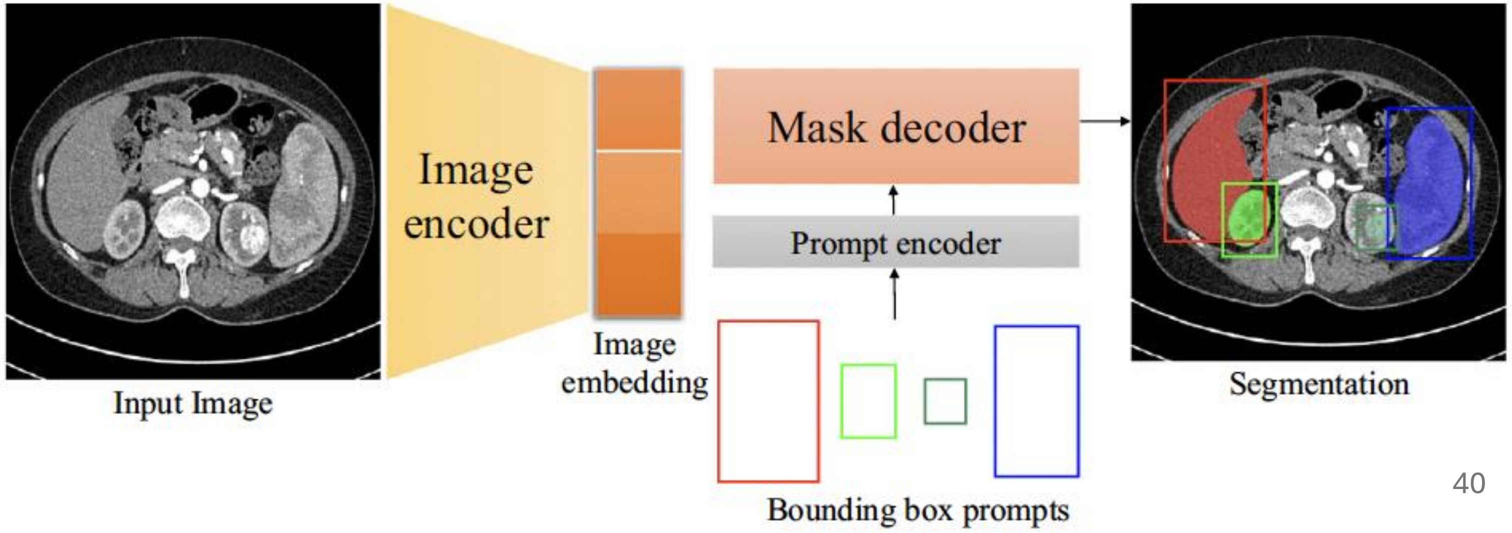
MedSAM

Interactive models (“SAM”)



Fine-tuning SAM on an huge amount of medical data

Pipeline

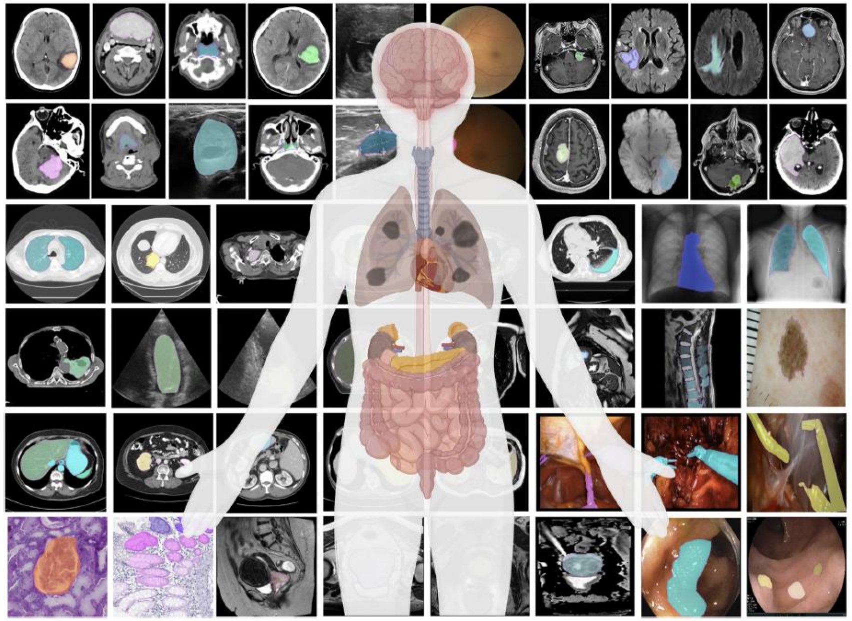


Ma et *al*. Segment Anything in Medical Images. Nat.Com.’24.

MedSAM

Learning / usage objectives.

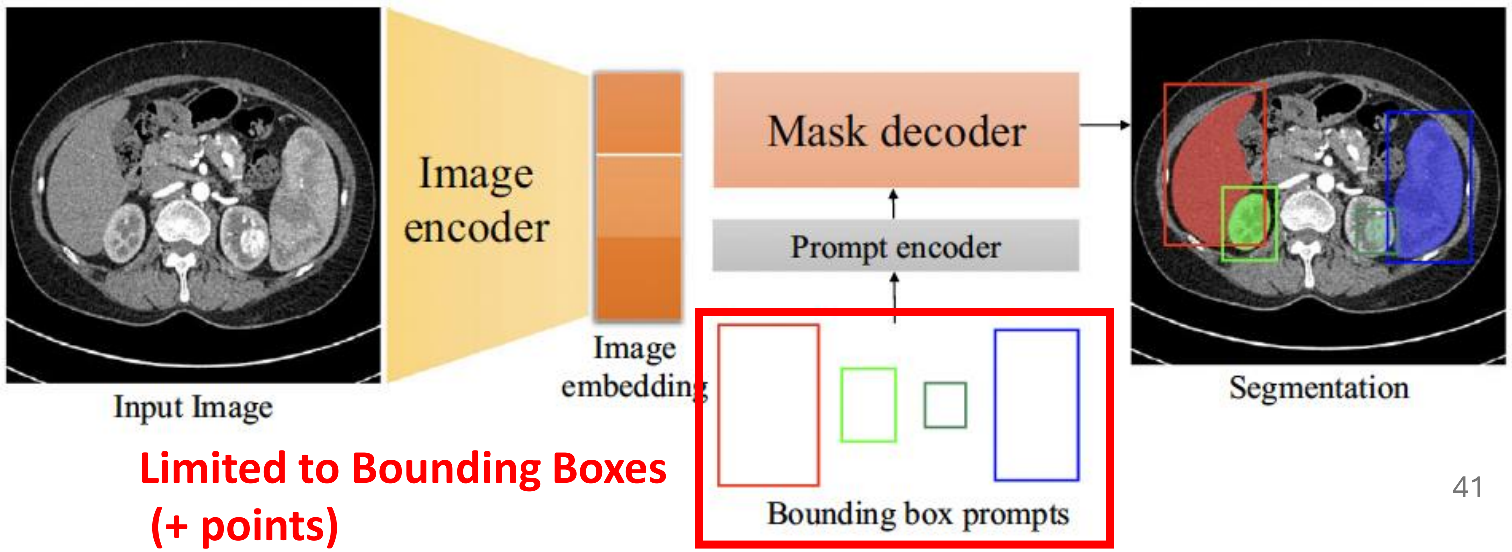
Interactive models ("SAM")



Trained on an huge amount of data

Pipeline

How good is the DSC of this bounding box?



Limited to Bounding Boxes (+ points)

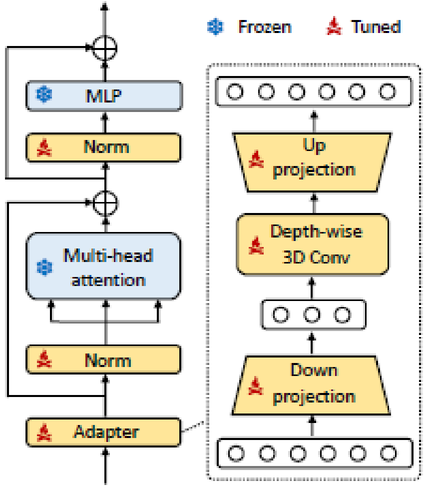
Ma et *al*. Segment Anything in Medical Images. Nat.Com.'24.

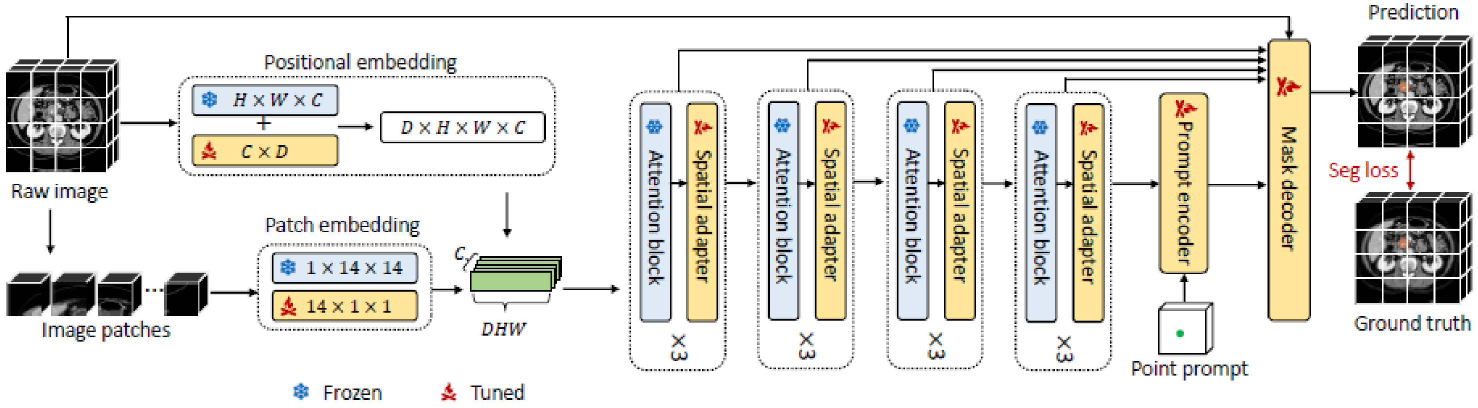
Learning / usage objectives.

3DSAM-Adapter

Interactive models ("SAM")

Fine-tuning SAM via Parameter-Efficient Fine-Tuning





Gong et *al*. 3DSAM-adapter: Holistic Adaptation of SAM from 2D to 3D for Promptable Medical Image Segmentation. MedIA'24.

Learning / usage objectives.

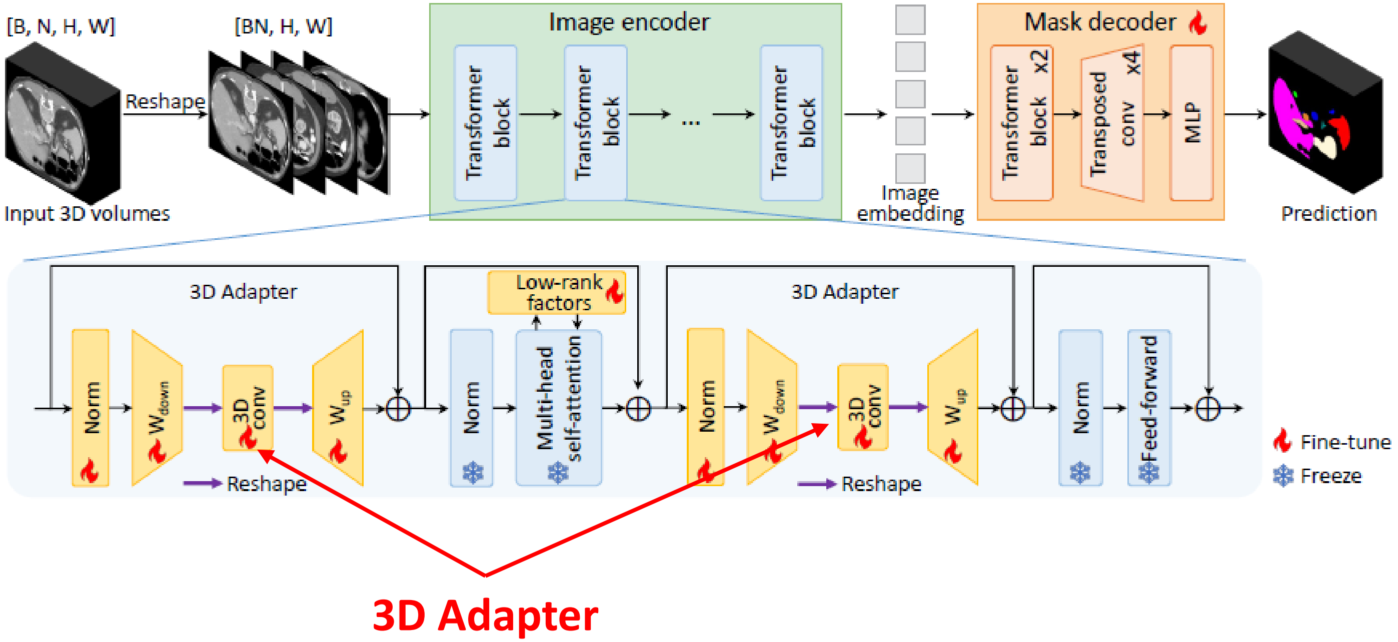
MA-SAM (3D)

Interactive models ("SAM")

Fine-tuning SAM 2D via Parameter-Efficient Fine-Tuning to 3D

→ Adapt for promptable version.

Methods	Dice ↑	NSD ↑
nnU-Net (Isensee et al., 2021)	41.6	62.5
3D UX-Net (Lee et al., 2023)	34.8	52.6
SwinUNETR (Tang et al., 2022b)	40.6	60.0
nnFormer (Zhou et al., 2023a)	36.5	54.0
3DSAM-adapter (automatic) (Gong et al., 2023)	30.2	45.4
3DSAM-adapter (10 pts/scan) (Gong et al., 2023)	57.5	79.6
MA-SAM (automatic)	40.2	59.1
MA-SAM (1 tight 3D bbx/scan)	**80.3**	**97.9**
MA-SAM (1 relaxed 3D bbx/scan)	74.7	97.1



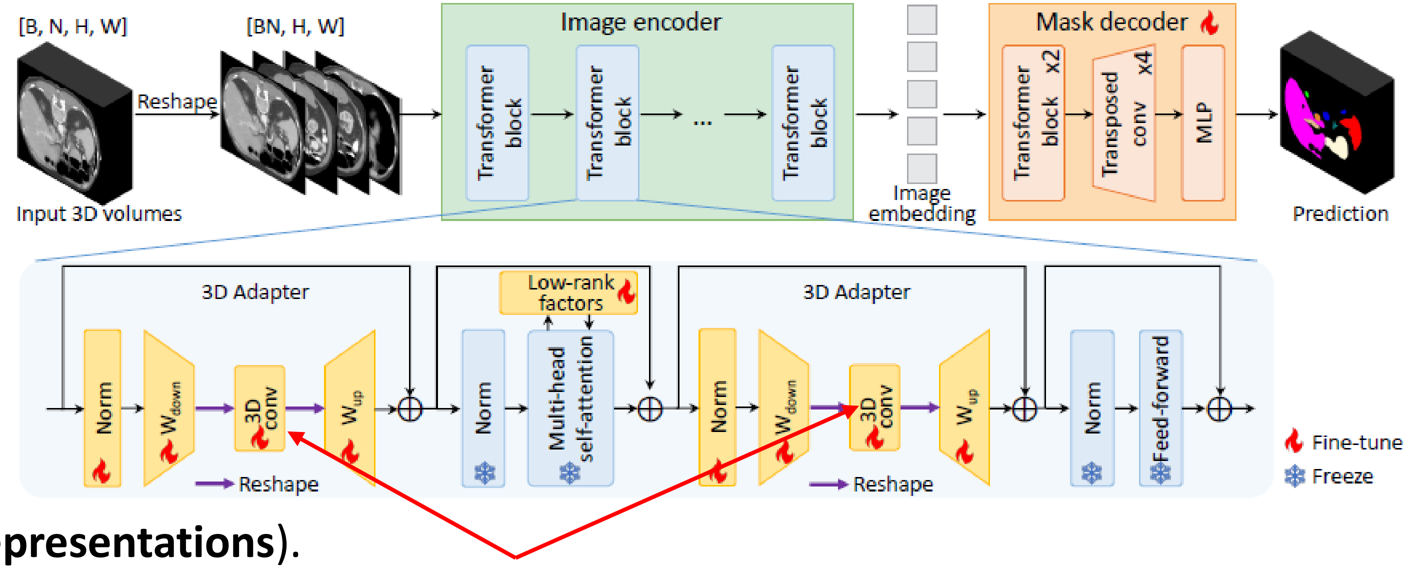
Chen et *al*. MA-SAM: Modality-agnostic SAM Adaptation for 3D Medical Image Segmentation. MedIA'24.

Learning / usage objectives.

MA-SAM (3D)

Interactive models (“SAM”)

Fine-tuning SAM 2D via Parameter-Efficient Fine-Tuning to 3D



3D Adapter

→ Adapt for zero-shot version (**SAM as pre-trained representations**).

Methods	Spleen	R.Kd	L.Kd	GB	Eso.	Liver	Stomach	Aorta	IVC	Veins	Pancreas	AG	Average
							Dice [%] ↑						
nnU-Net (Isensee et al., 2021)	**97.0**	**95.3**	95.3	63.5	77.5	**97.4**	89.1	90.1	**88.5**	79.0	**87.1**	**75.2**	86.3
3D UX-Net (Lee et al., 2023)	94.6	94.2	94.3	59.3	72.2	96.4	73.4	87.2	84.9	72.2	80.9	67.1	81.4
SwinUNETR (Tang et al., 2022b)	95.6	94.2	94.3	63.6	75.5	96.6	79.2	89.9	83.7	75.0	82.2	67.3	83.1
nnFormer (Zhou et al., 2023a)	93.5	94.9	95.0	64.1	79.5	96.8	90.1	89.7	85.9	77.8	85.6	73.9	85.6
SAMed_h (Zhang and Liu, 2023)	95.3	92.1	92.9	62.1	75.3	96.4	90.2	87.6	79.8	74.2	77.9	61.0	82.1
MA-SAM (Ours)	96.7	95.1	**95.4**	**68.2**	**82.1**	96.9	**92.8**	**91.1**	87.5	**79.8**	86.6	73.9	**87.2**

+0.9%

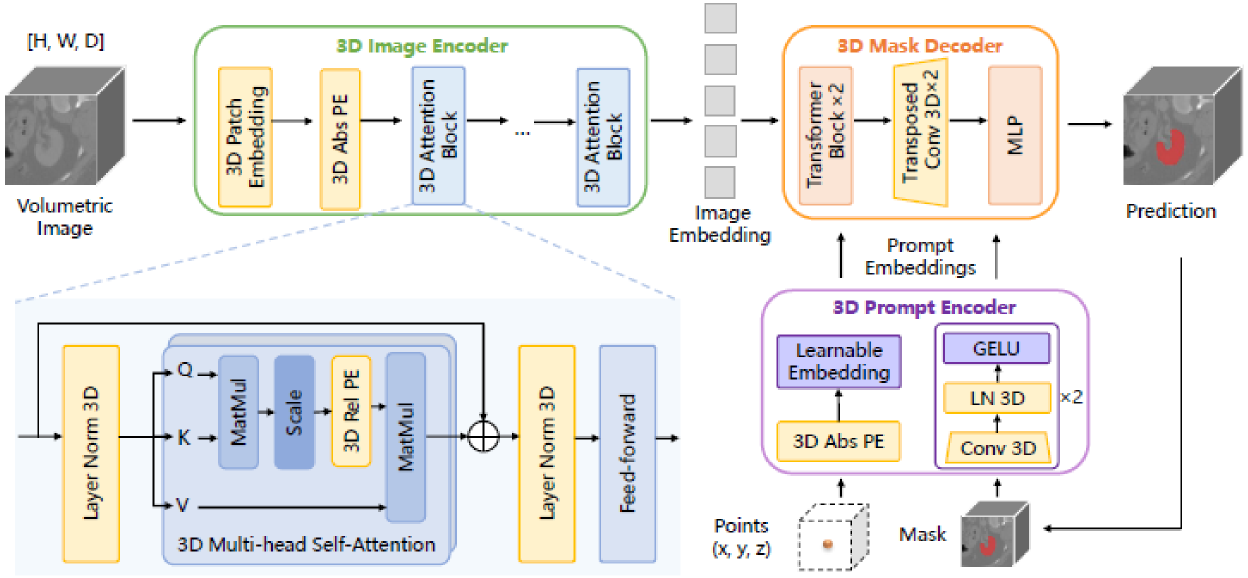
Chen et *al*. MA-SAM: Modality-agnostic SAM Adaptation for 3D Medical Image Segmentation. MedIA’24.

Learning / usage objectives.

Med-SAM3D

Training a 3D SAM with Medical data from Scratch

Interactive models ("SAM")



Model	Prompt	Inference Time (s)	Dice (%)		
			Seen	Unseen	Overall
SAM	N pts	$N(\tau+0.13)$	16.79	11.73	16.15
SAM-Med2D	N pts	$N(\tau+0.04)$	38.91	22.55	36.83
SAM-Med3D	1 pt	τ+2	81.98	37.02	76.27
SAM	$3N$ pts	$3N(\tau+0.19)$	34.61	15.94	32.24
SAM-Med2D	$3N$ pts	$3N(\tau+0.07)$	51.46	29.70	48.70
SAM-Med3D	3 pts	3τ+3	84.14	43.80	79.02
SAM	$5N$ pts	$5N(\tau+0.25)$	49.39	21.86	45.89
SAM-Med2D	$5N$ pts	$5N(\tau+0.10)$	51.89	30.41	49.17
SAM-Med3D	5 pts	5τ+4	84.62	46.26	79.75
SAM-Med3D	10 pts	10τ+6	85.19	49.92	80.71

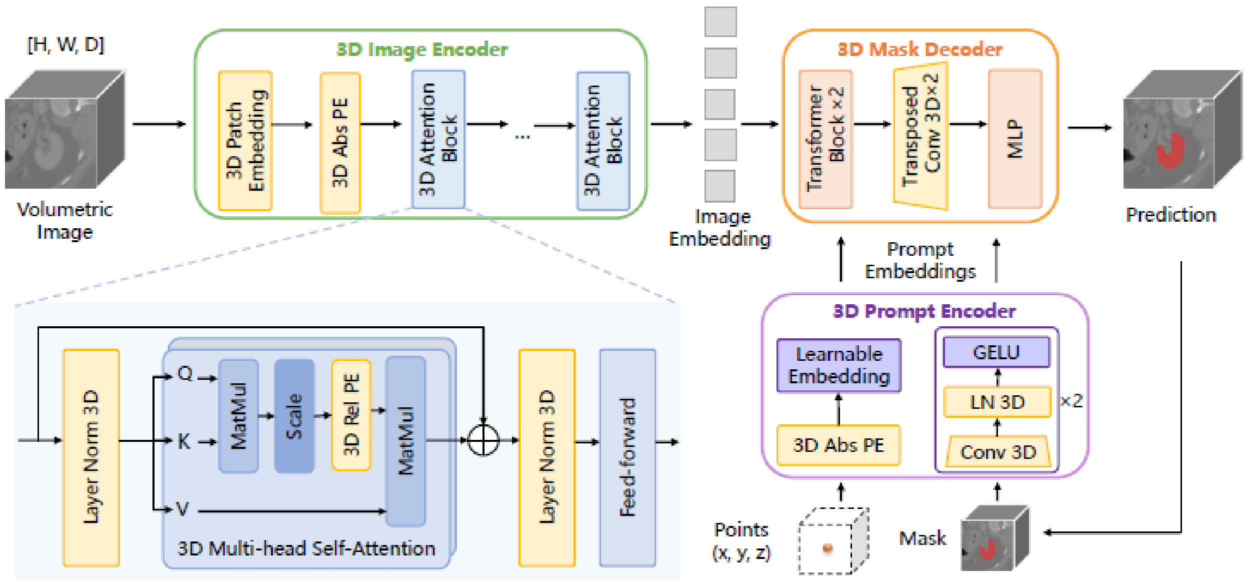
Wang et *al*. SAM-Med3D: Towards General-Purpose Segmentation Models for Volumetric Medical Images. ArXiv'24.

Learning / usage objectives.

Med-SAM3D

Training a 3D SAM with Medical data from Scratch

Interactive models ("SAM")



1 point for each N slices

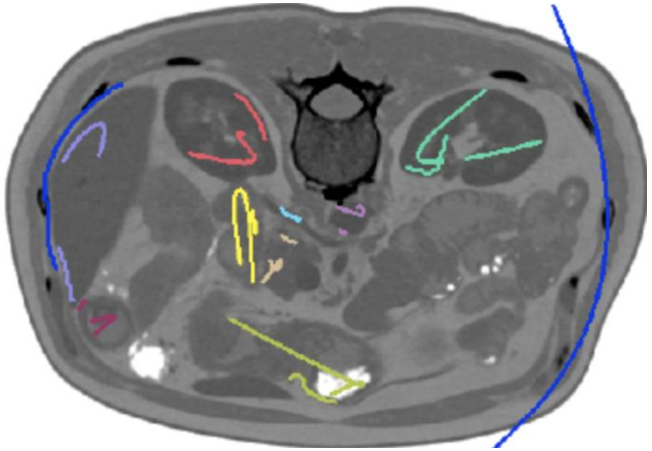
Model	Prompt	Inference Time (s)	Dice (%) Seen	Unseen	Overall
SAM	N pts	$N(\tau+0.13)$	16.79	11.73	16.15
SAM-Med2D	N pts	$N(\tau+0.04)$	38.91	22.55	36.83
SAM-Med3D	1 pt	τ+2	81.98	37.02	76.27
SAM	$3N$ pts	$3N(\tau+0.19)$	34.61	15.94	32.24
SAM-Med2D	$3N$ pts	$3N(\tau+0.07)$	51.46	29.70	48.70
SAM-Med3D	3 pts	3τ+3	84.14	43.80	79.02
SAM	$5N$ pts	$5N(\tau+0.25)$	49.39	21.86	45.89
SAM-Med2D	$5N$ pts	$5N(\tau+0.10)$	51.89	30.41	49.17
SAM-Med3D	5 pts	5τ+4	84.62	46.26	79.75
SAM-Med3D	10 pts	10τ+6	85.19	49.92	80.71

Improved over 2D version

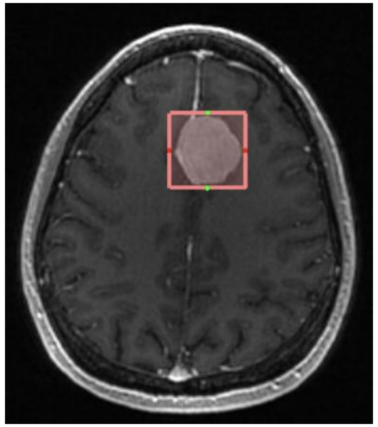
Wang et *al*. SAM-Med3D: Towards General-Purpose Segmentation Models for Volumetric Medical Images. ArXiv'24.

Learning / usage objectives.

Interactive models ("SAM")



SAM is promptable
(i.e., requires user interaction
per EACH test image)



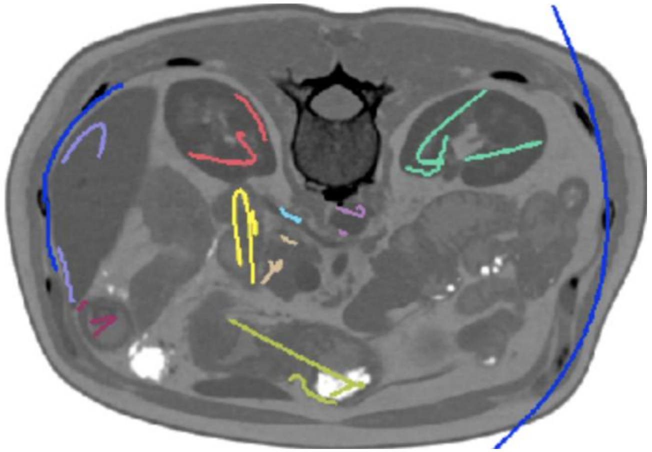
SAM only handles
binary segmentation
(one class at a time)

Wang et *al*. SAM-Med3D: Towards General-Purpose Segmentation
Models for Volumetric Medical Images. ArXiv'24.

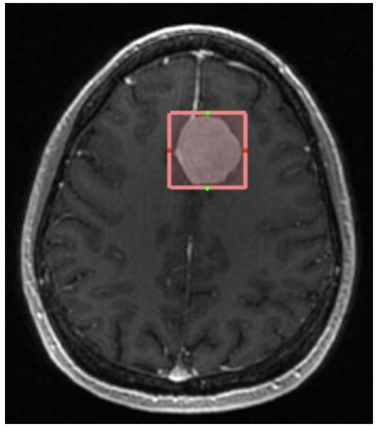
Med-SAM3D

Learning / usage objectives.

Interactive models ("SAM")



SAM is promptable
(i.e., requires user interaction
per EACH test image)



SAM only handles
binary segmentation
(one class at a time)

Dataset	**Modality**	**Task-specific** UNETR [11]	nnU-Net [16]	**General-purpose** SAM-Med2D [6] (N pts)	SegVol [8] (pt+text)	Ours (1 pt)	Ours (10 pts)
Totalsegmentator [36]	CT	75.05	85.22	38.26	-	84.68	**87.59**
KiTS21 [12]	CT	70.75	75.32	68.74	-	72.06	**75.37**
AMOS-CT [17]	CT	78.33	**88.87**	49.61	-	79.94	83.99
AMOS-MR [17]	MR	76.29	**86.92**	45.53	-	75.41	81.13
BTCV* [19]	CT	78.99	81.92	50.05	73.81	79.17	**83.01**
TDSC-ABUS23* [33]	US*	-	45.08	49.39	-	36.08	**54.35**

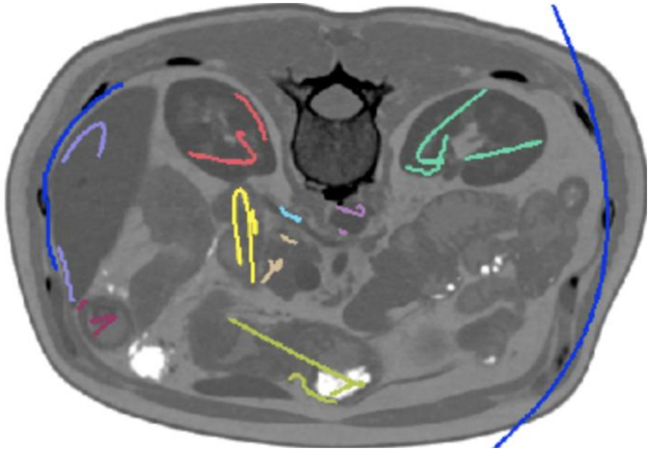
SAM yields sometimes
lower results to task-
specific models

Wang et *al*. SAM-Med3D: Towards General-Purpose Segmentation Models for Volumetric Medical Images. ArXiv'24.

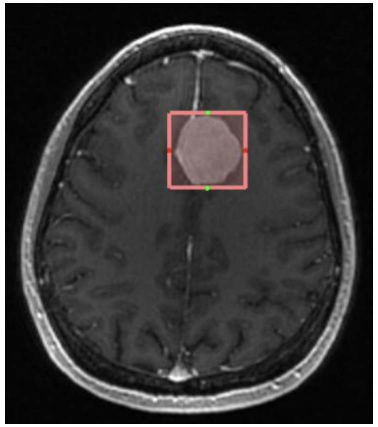
Learning / usage objectives.

Med-SAM3D

Interactive models ("SAM")



SAM is promptable
(i.e., requires user interaction
per EACH test image)



SAM only handles
binary segmentation
(one class at a time)

Dataset	Modality	Task-specific		General-purpose			
		UNETR [11]	nnU-Net [16]	SAM-Med2D [6] (N pts)	SegVol [8] (pt+text)	Ours (1 pt)	Ours (10 pts)
Totalsegmentator [36]	CT	75.05	85.22	38.26	-	84.68	**87.59**
KiTS21 [12]	CT	70.75	75.32	68.74	-	72.06	**75.37**
AMOS-CT [17]	CT	78.33	**88.87**	49.61	-	79.94	83.99
AMOS-MR [17]	MR	76.29	**86.92**	45.53	-	75.41	81.13
BTCV* [19]	CT	78.99	81.92	50.05	73.81	79.17	**83.01**
TDSC-ABUS23* [33]	US*	-	45.08	49.39	-	36.08	**54.35**

SAM yields sometimes
lower results to task-
specific models

Wang et *al*. SAM-Med3D: Towards General-Purpose Segmentation Models for Volumetric Medical Images. ArXiv'24.

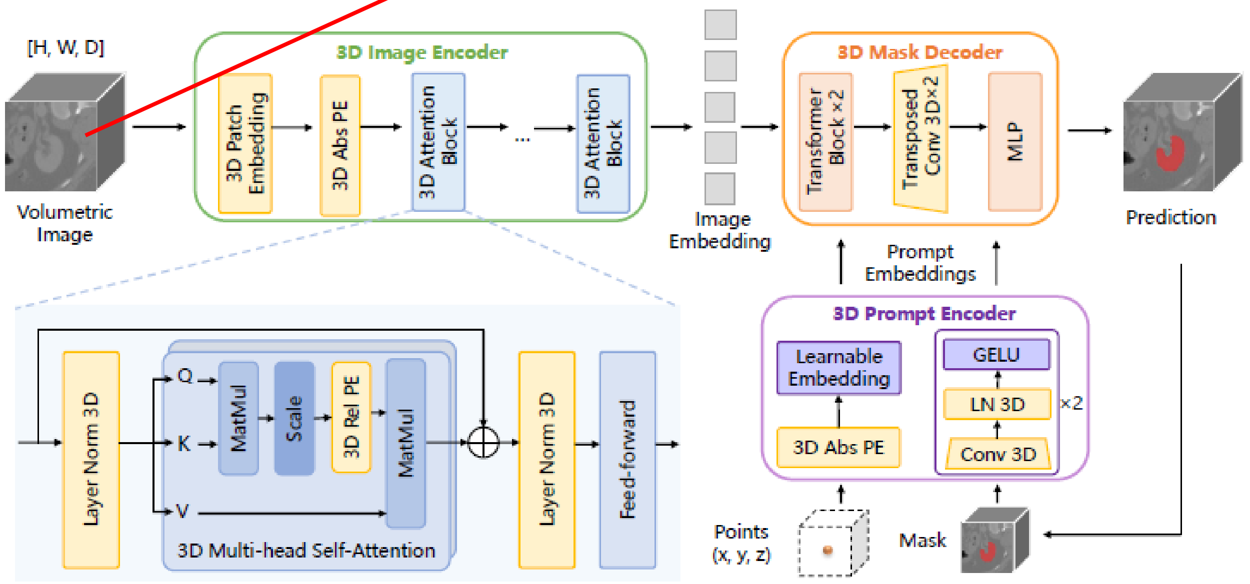
Learning / usage objectives.

Med-SAM3D

Other Details

Interactive models ("SAM")

Pre-computed ROI in whole-body scans



Model	Prompt	Inference Time (s)	Dice (%) Seen	Unseen	Overall
SAM	N pts	$N(\tau+0.13)$	16.79	11.73	16.15
SAM-Med2D	N pts	$N(\tau+0.04)$	38.91	22.55	36.83
SAM-Med3D	1 pt	τ+2	81.98	37.02	76.27
SAM	$3N$ pts	$3N(\tau+0.19)$	34.61	15.94	32.24
SAM-Med2D	$3N$ pts	$3N(\tau+0.07)$	51.46	29.70	48.70
SAM-Med3D	3 pts	3τ+3	84.14	43.80	79.02
SAM	$5N$ pts	$5N(\tau+0.25)$	49.39	21.86	45.89
SAM-Med2D	$5N$ pts	$5N(\tau+0.10)$	51.89	30.41	49.17
SAM-Med3D	5 pts	5τ+4	84.62	46.26	79.75
SAM-Med3D	10 pts	10τ+6	85.19	49.92	80.71

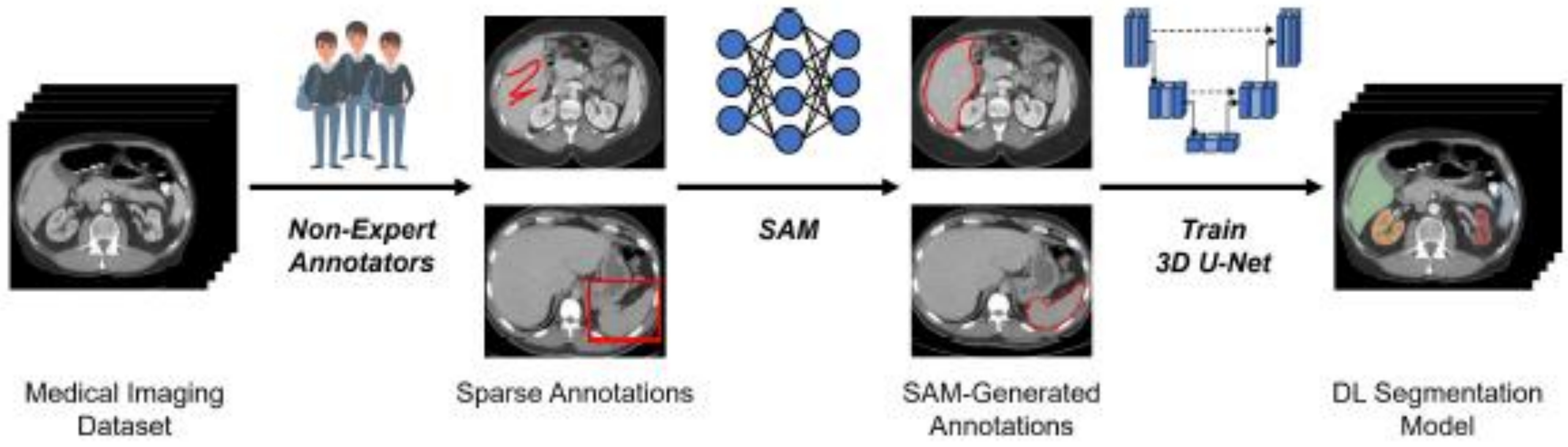
Iterative random points over the error region (explicit access to GT)

Wang et *al*. SAM-Med3D: Towards General-Purpose Segmentation Models for Volumetric Medical Images. ArXiv'24.

Learning / usage objectives.

Interactive models (“SAM”)

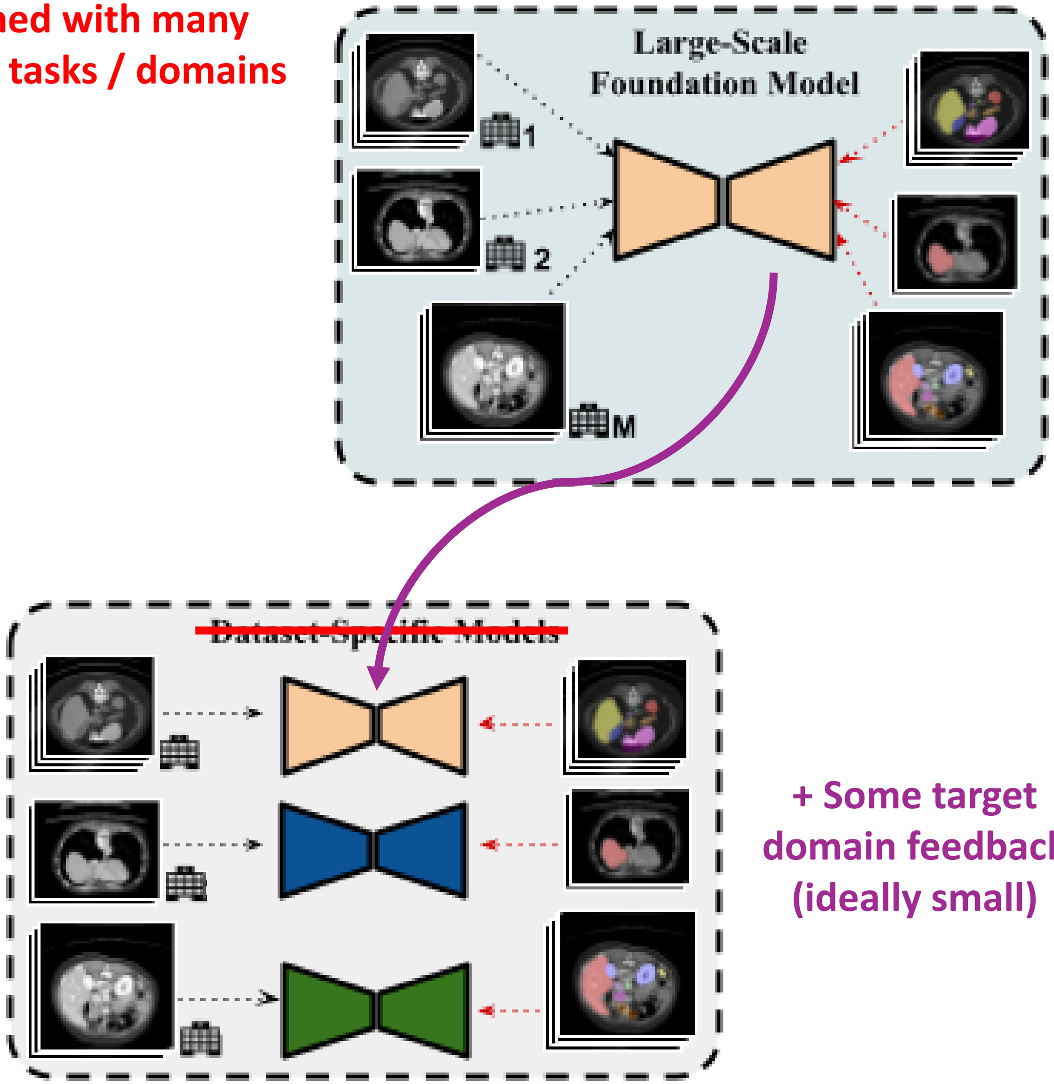
Applications in Active Learning / Annotations



Kulkarni et *al*. Anytime, Anywhere, Anyone: Investigating the Feasibility of SAM for Crowd-Sourcing Medical Image Annotations. MIDL’24.

Foundation models for medical image segmentation

Trained with many data / tasks / domains



+ Some target domain feedback (ideally small)

Organizing the mess!

1. Types of foundation models: a data perspective.
   A. Generalist vs. Specialized
   B. 2D *vs*. 3D
   C. Multimodal *vs*. Unimodal
2. Learning/Usage Objectives
   A. Zero-shot / Transfer Learning
   B. In-Context Learning
   C. Interactive Models (“SAM”)
3. **Zero-shot / Adaptation-oriented (3D data)**
   A. How to pre-train?
   B. How useful are foundation models? Limitations on the adaptation stage
   C. Few-shot Parameter-Efficient Fine-tuning

Learning / usage objectives.

Zero-shot / Transfer Learning

Med3D('19)

CLIP-Driven

MultiTalent

HERMES

UniSeg

FSEFT

SuPreM

ImageNet Philosophy

Zero-shot predictions to *base tasks*

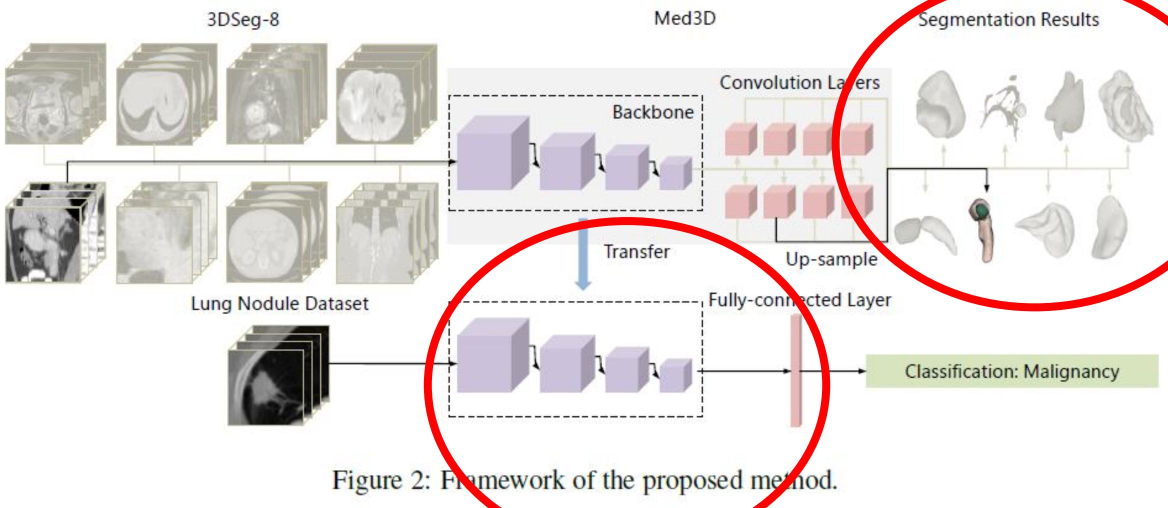
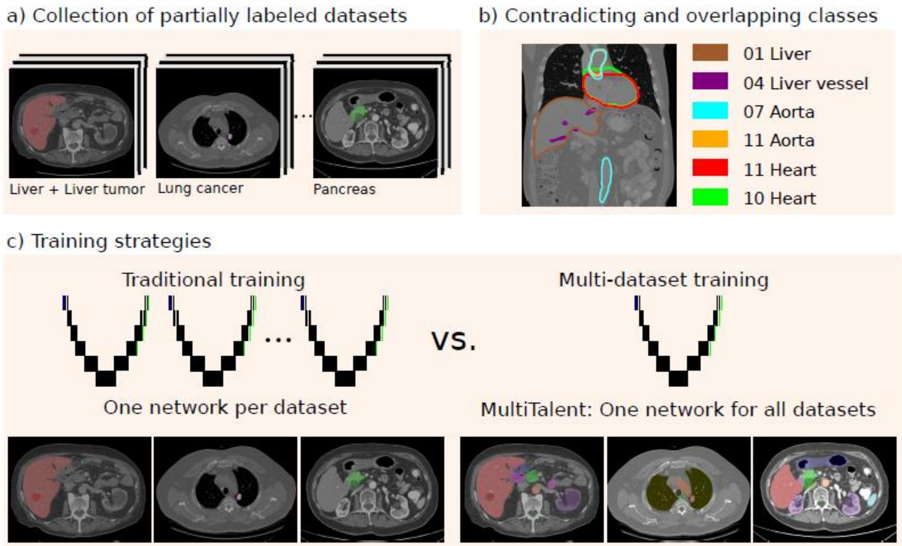


Figure 2: Framework of the proposed method.

Chen et *al*. Med3D: Transfer Learning for 3D Medical Image Analysis. ArXiv'19.

Fine-tuning to *novel domains/tasks*



Ulrich et *al*. MultiTalent: A Multi-Dataset Approach to Medical Image Segmentation. MICCAI'23.

Zero-shot /Adaptation Oriented (3D Data)

Med3D('19)

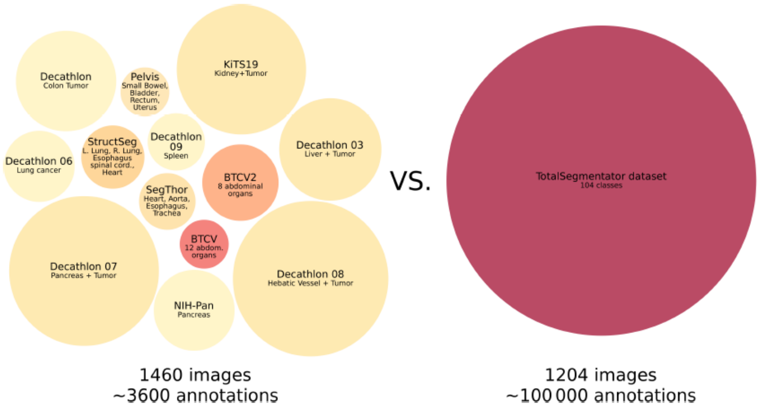
CLIP-Driven

MultiTalent

UniSeg

SuPreM

Why volumetric (and mostly CT)?



Ulrich et *al*. MultiTalent: A Multi-Dataset Approach to Medical Image Segmentation. MICCAI'23.

Datasets	# Targets	# Scans	Annotated Organs or Tumors
1. Pancreas-CT [62]	1	82	Pancreas
2. LiTS [3]	2	201	Liver, Liver Tumor*
3. KiTS [25]	2	300	Kidney, Kidney Tumor*
4. AbdomenCT-1K [45]	4	1,000	Spleen, Kidney, Liver, Pancreas
5. CT-ORG [60]	4	140	Lung, Liver, Kidneys and Bladder
6. CHAOS [73]	4	40	Liver, Left Kidney, Right Kidney, Spl
7-11. MSD CT Tasks [1]	9	947	Spl, Liver and Tumor*, Lung Tumor*, Colon Tumor*, Pan and Tumor*, Hepatic Vessel and Tumor*
12. BTCV [37]	13	50	Spl, RKid, LKid, Gall, Eso, Liv, Sto, Aor, IVC, R&SVeins, Pan, RAG, LAG
13. AMOS22 [32]	15	500	Spl, RKid, LKid, Gall, Eso, Liv, Sto, Aor, IVC, Pan, RAG, LAG, Duo, Bla, Pro/UTE
14. WORD [44]	16	150	Spl, RKid, LKid, Gall, Eso, Liv, Sto, Pan, RAG, Duo, Col, Int, Rec, Bla, LFH, RFH
15. 3D-IRCADb [67]	13	20	Liv, Liv Cyst, RLung, LLung, Venous, PVein, Aor, Spl, RKid, LKid, Gall, IVC
16. TotalSegmentator [79]	104	1,024	Clavicula, Humerus, Scapula, Rib 1-12, Vertebrae C1-7, Vertebrae T1-9, Vertebrae L1-5, Hip, Sacrum, Femur, Aorta, Pulmonary Artery, Right Ventricle, Right Atrium, Left Atrium, Left Ventricle, Myocardium, PVein, SVein, IVC, Iliac Artery, Iliac Vena, Brain, Trachea, Lung Upper Lobe, Lung Middle Lobe, Lung Lower Lobe, AG, Spl, Liv, Gall, Pan, Kid, Eso, Sto, Duo, Small Bowel, Colon, Bla, Autochthon, Iliopsoas, Gluteus Minimus, Gluteus Medius, Gluteus Maximus
17. JHH (*private*)	21	5,038	Aor, AG, CBD, Celiac AA, Colon, duo, Gall, IVC, Lkid, RKid, Liv, Pan, Pan Duct, SMA, Small bowel, Spl, Sto, Veins, Kid LtRV, Kid RtRV, CBD Stent, PDAC*, PanNET*, Pancreatic Cyst*

Liu et *al*. CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection. ICCV'23.

Zero-shot /Adaptation Oriented (3D Data)

Med3D('19)

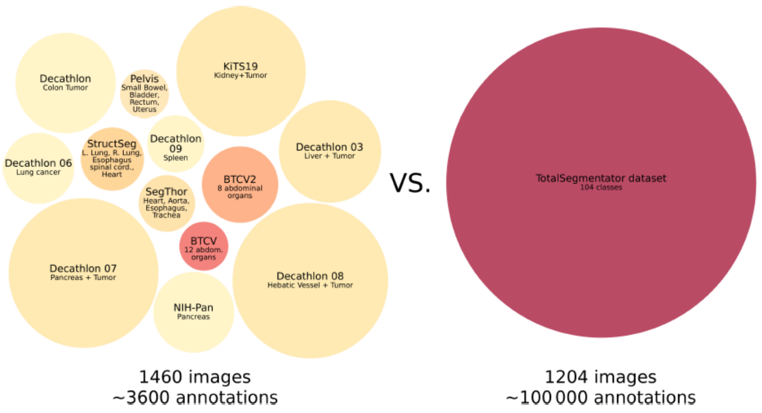
CLIP-Driven

MultiTalent

UniSeg

SuPreM

Why volumetric (and mostly CT)?



Ulrich et *al*. MultiTalent: A Multi-Dataset Approach to Medical Image Segmentation. MICCAI'23.

→ A good number of annotated scans publicly available. **(current models are pre-trained with 2K CTs)**

→ Anatomical morphology is natural 3D.

→ Labeling at voxel level is tremendously costly for practitioners.

(10 min per structure according to TotalSegmentator).

→ Enormous potential of FMs to address inter-center, inter-scan and demographics variabilities.

Datasets	# Targets	# Scans	Annotated Organs or Tumors
1. Pancreas-CT [62]	1	82	Pancreas
2. LiTS [3]	2	201	Liver, Liver Tumor*
3. KiTS [25]	2	300	Kidney, Kidney Tumor*
4. AbdomenCT-1K [45]	4	1,000	Spleen, Kidney, Liver, Pancreas
5. CT-ORG [60]	4	140	Lung, Liver, Kidneys and Bladder
6. CHAOS [73]	4	40	Liver, Left Kidney, Right Kidney, Spl
7-11. MSD CT Tasks [1]	9	947	Spl, Liver and Tumor*, Lung Tumor*, Colon Tumor*, Pan and Tumor*, Hepatic Vessel and Tumor*
12. BTCV [37]	13	50	Spl, RKid, LKid, Gall, Eso, Liv, Sto, Aor, IVC, R&SVeins, Pan, RAG, LAG
13. AMOS22 [32]	15	500	Spl, RKid, LKid, Gall, Eso, Liv, Sto, Aor, IVC, Pan, RAG, LAG, Duo, Bla, Pro/UTE
14. WORD [44]	16	150	Spl, RKid, LKid, Gall, Eso, Liv, Sto, Pan, RAG, Duo, Col, Int, Rec, Bla, LFH, RFH
15. 3D-IRCADb [67]	13	20	Liv, Liv Cyst, RLung, LLung, Venous, PVein, Aor, Spl, RKid, LKid, Gall, IVC
16. TotalSegmentator [79]	104	1,024	Clavicula, Humerus, Scapula, Rib 1-12, Vertebrae C1-7, Vertebrae T1-9, Vertebrae L1-5, Hip, Sacrum, Femur, Aorta, Pulmonary Artery, Right Ventricle, Right Atrium, Left Atrium, Left Ventricle, Myocardium, PVein, SVein, IVC, Iliac Artery, Iliac Vena, Brain, Trachea, Lung Upper Lobe, Lung Middle Lobe, Lung Lower Lobe, AG, Spl, Liv, Gall, Pan, Kid, Eso, Sto, Duo, Small Bowel, Colon, Bla, Autochthon, Iliopsoas, Gluteus Minimus, Gluteus Medius, Gluteus Maximus
17. JHH (*private*)	21	5,038	Aor, AG, CBD, Celiac AA, Colon, duo, Gall, IVC, Lkid, RKid, Liv, Pan, Pan Duct, SMA, Small bowel, Spl, Sto, Veins, Kid LtRV, Kid RtRV, CBD Stent, PDAC*, PanNET*, Pancreatic Cyst*

Liu et *al*. CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection. ICCV'23.

Zero-shot /Adaptation Oriented (3D Data)

Med3D('19)

CLIP-Driven

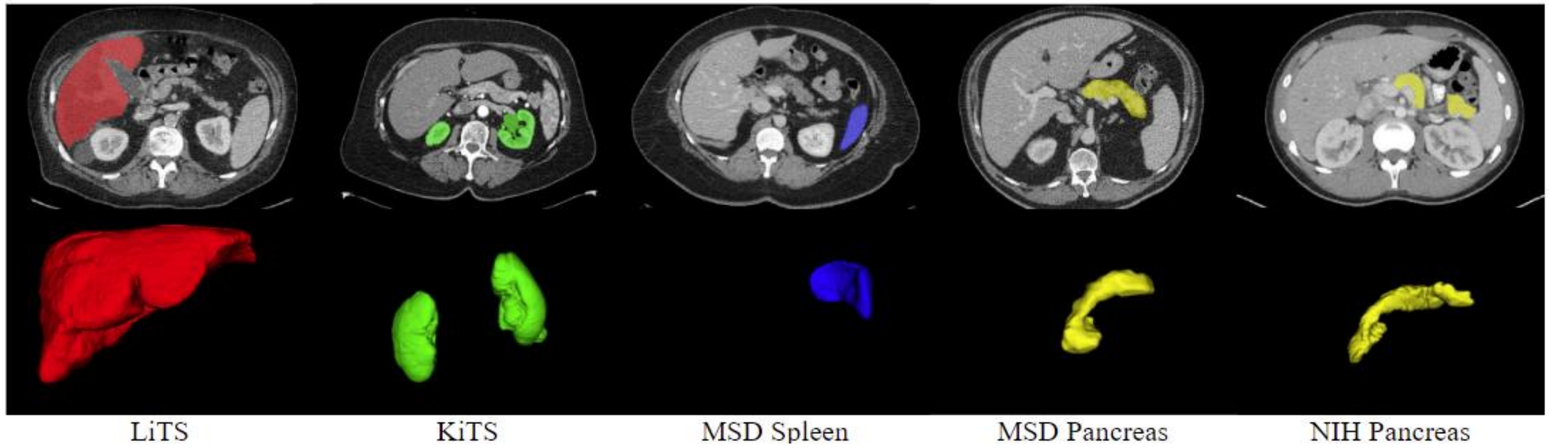
MultiTalent

UniSeg

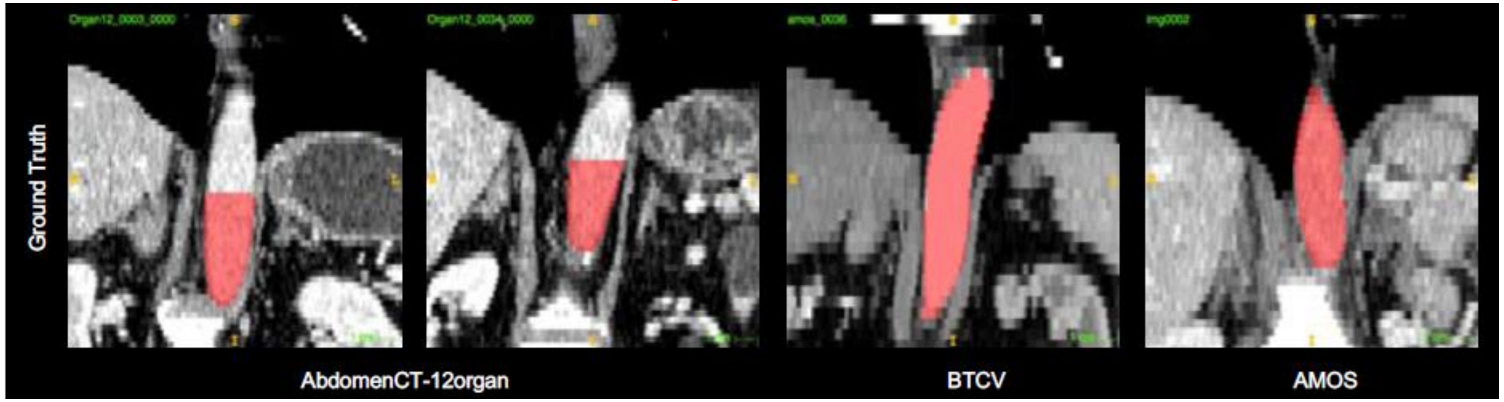
SuPreM

Challenges of Dataset Assembling

Partially-labeled datasets



Inconsistent annotation protocols

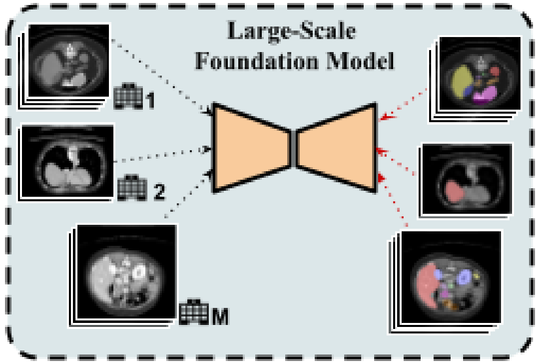


Liu et *al*. CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection. ICCV'23.

Zero-shot /Adaptation Oriented (3D Data)

MultiTalent

FSEFT



How to pre-train? Standard

Assembly Dataset with
Partial Labels

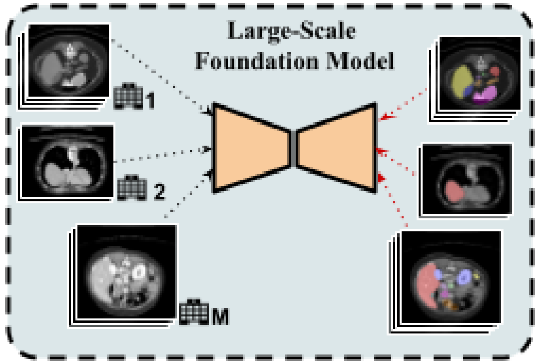
$$\mathcal{D}_T = \{(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{w}_n)\}_{n=1}^{N}$$

Silva-Rodríguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MICCAI W-MedAGI'23 (Extension Under Review).

Zero-shot / Adaptation Oriented (3D Data)

MultiTalent

FSEFT



How to pre-train? Standard

Total Number of Categories

$$w^c = [0, 1, 1, 0, 0, 0, 1, 0, 0]$$

Assembly Dataset with Partial Labels

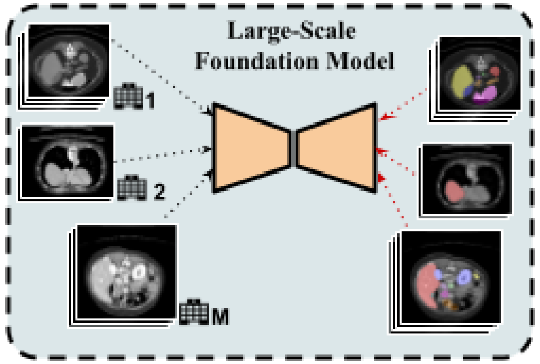
$$\mathcal{D}_T = \{(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{w}_n)\}_{n=1}^{N}$$

Silva-Rodríguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MICCAI W-MedAGI'23 (Extension Under Review).

Zero-shot /Adaptation Oriented (3D Data)

MultiTalent

FSEFT



How to pre-train? Standard

Annotated on its dataset

$$w^c = [0, 1, 1, 0, 0, 0, 1, 0, 0]$$

Assembly Dataset with Partial Labels

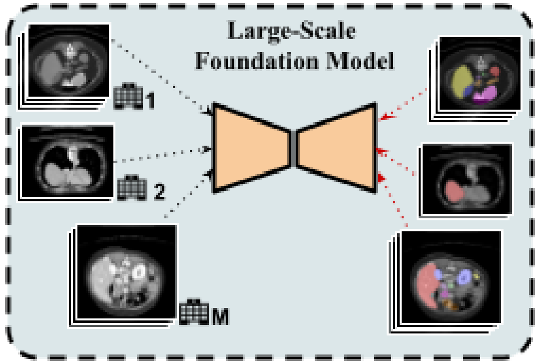
$$\mathcal{D}_T = \{(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{w}_n)\}_{n=1}^N$$

Silva-Rodríguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MICCAI W-MedAGI'23 (Extension Under Review).

Zero-shot / Adaptation Oriented (3D Data)

MultiTalent

FSEFT



How to pre-train? Standard

NOT annotated on its dataset

$$w^c = [0, 1, 1, 0, 0, 0, 1, 0, 0]$$

Assembly Dataset with Partial Labels

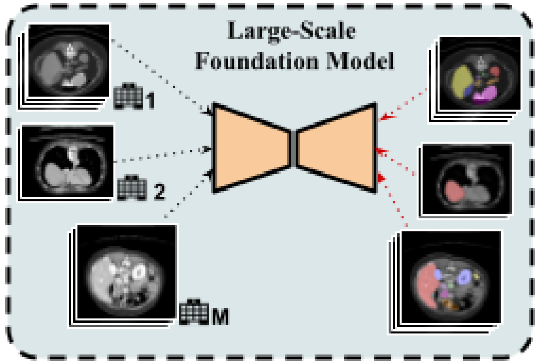
$$\mathcal{D}_T = \{(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{w}_n)\}_{n=1}^{N}$$

Silva-Rodríguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MICCAI W-MedAGI'23 (Extension Under Review).

Zero-shot / Adaptation Oriented (3D Data)

MultiTalent

FSEFT



How to pre-train? Standard

Assembly Dataset with
Partial Labels

$$\mathcal{D}_T = \{(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{w}_n)\}_{n=1}^{N}$$

1. Forward Encoder-Decoder

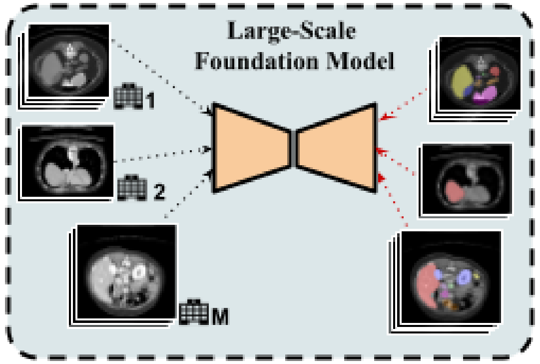
$$\mathbf{Z}_n = \theta_f(\mathbf{X}_n)$$

Silva-Rodríguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MICCAI W-MedAGI'23 (Extension Under Review).

Zero-shot /Adaptation Oriented (3D Data)

MultiTalent

FSEFT



How to pre-train? Standard

Assembly Dataset with
Partial Labels

$$\mathcal{D}_T = \{(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{w}_n)\}_{n=1}^{N}$$

1. Forward Encoder-Decoder

$$\mathbf{Z}_n = \theta_f(\mathbf{X}_n)$$

2. Forward Classifier + Sigmoid activation

$$\hat{\mathbf{Y}}_n = \sigma(\theta_c(\mathbf{Z}_n))$$

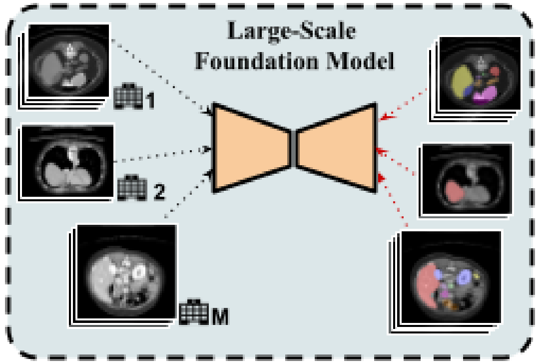
Disentangle prediction
for each task
(softmax might affect not-
annotated categories)

Silva-Rodríguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MICCAI W-MedAGI'23 (Extension Under Review).

Zero-shot /Adaptation Oriented (3D Data)

MultiTalent

FSEFT



How to pre-train? Standard

Assembly Dataset with
Partial Labels

$$\mathcal{D}_T = \{(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{w}_n)\}_{n=1}^{N}$$

1. Forward Encoder-Decoder

$$\mathbf{Z}_n = \theta_f(\mathbf{X}_n)$$

2. Forward Classifier + Sigmoid activation

$$\hat{\mathbf{Y}}_n = \sigma(\theta_c(\mathbf{Z}_n))$$

3. Compute any masked segmentation loss, and update

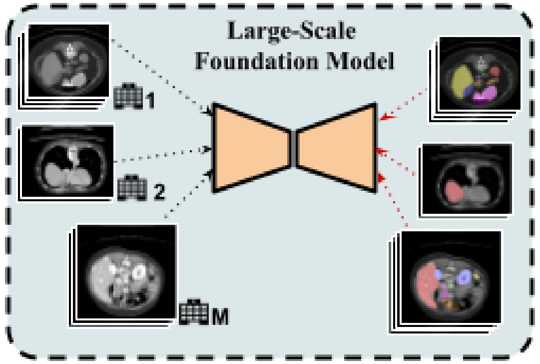
$$\boxed{\min_{\theta_f,\theta_c} \quad \frac{1}{\sum_k \mathbf{w}_{n,k}} \sum_k \mathbf{w}_{n,k} \mathcal{L}_{SEG}(\mathbf{Y}_{n,k}, \hat{\mathbf{Y}}_{n,k}), \quad n = 1, \ldots, N}$$

Silva-Rodríguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MICCAI W-MedAGI'23 (Extension Under Review).

Zero-shot /Adaptation Oriented (3D Data)

MultiTalent

FSEFT



How to pre-train? Standard

Assembly Dataset with
Partial Labels

$$\mathcal{D}_T = \{(\mathbf{X}_n, \mathbf{Y}_n, \mathbf{w}_n)\}_{n=1}^{N}$$

1. Forward Encoder-Decoder

$$\mathbf{Z}_n = \theta_f(\mathbf{X}_n)$$

2. Forward Classifier + Sigmoid activation

$$\hat{\mathbf{Y}}_n = \sigma(\theta_c(\mathbf{Z}_n))$$

3. Compute any masked segmentation loss, and update

$$L = \sum_c \left(\mathbb{1}_c^{(k)} \frac{1}{I} \sum_z BCE(\hat{y}_{z,c}^{(k)},\ y_{z,c}^{(k)}) - \frac{2 \sum_z \mathbb{1}_c^{(k)}\ \hat{y}_{z,c}^{(k)}\ y_{z,c}^{(k)}}{\sum_z \mathbb{1}_c^{(k)}\ \hat{y}_{z,c}^{(k)} + \sum_z \mathbb{1}_c^{(k)}\ y_{z,c}^{(k)}} \right)$$

Ulrich et *al*. MultiTalent: A Multi-Dataset Approach to Medical Image Segmentation. MICCAI'23.

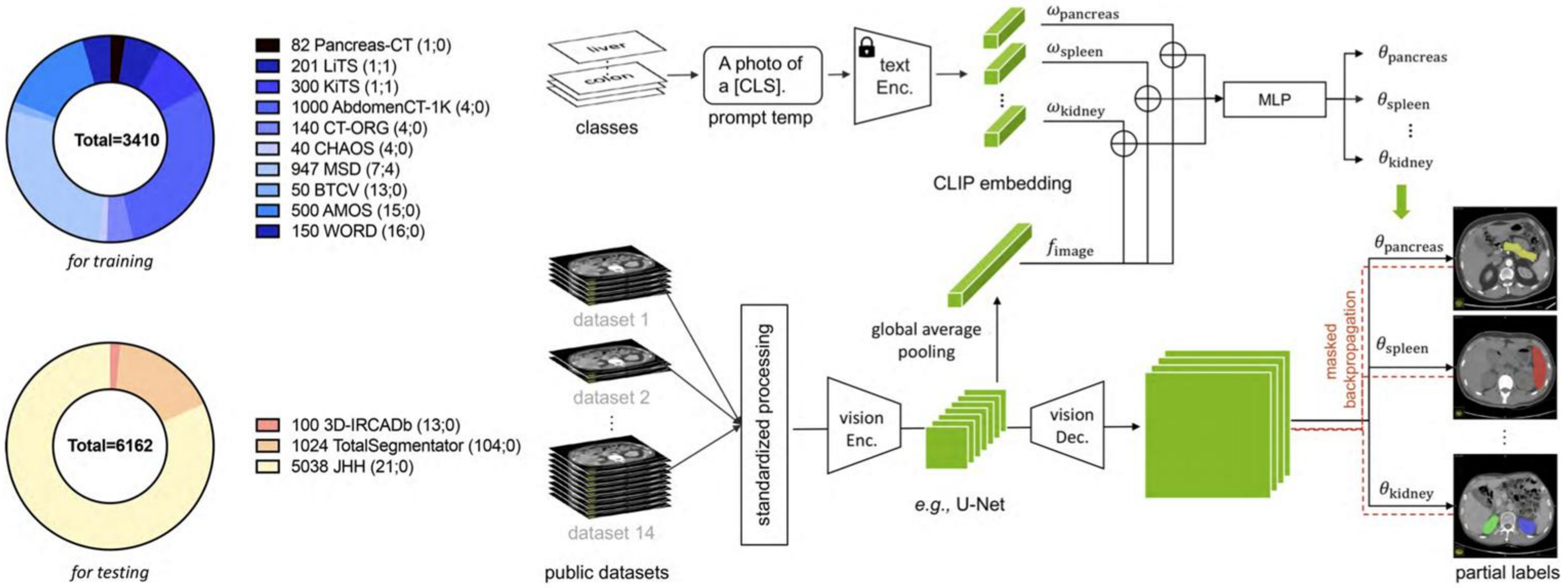
Zero-shot /Adaptation Oriented (3D Data)

CLIP-Driven

SuPreM

How to pre-train? CLIP-Driven

Main idea



Liu et *al*. CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection. ICCV'23.

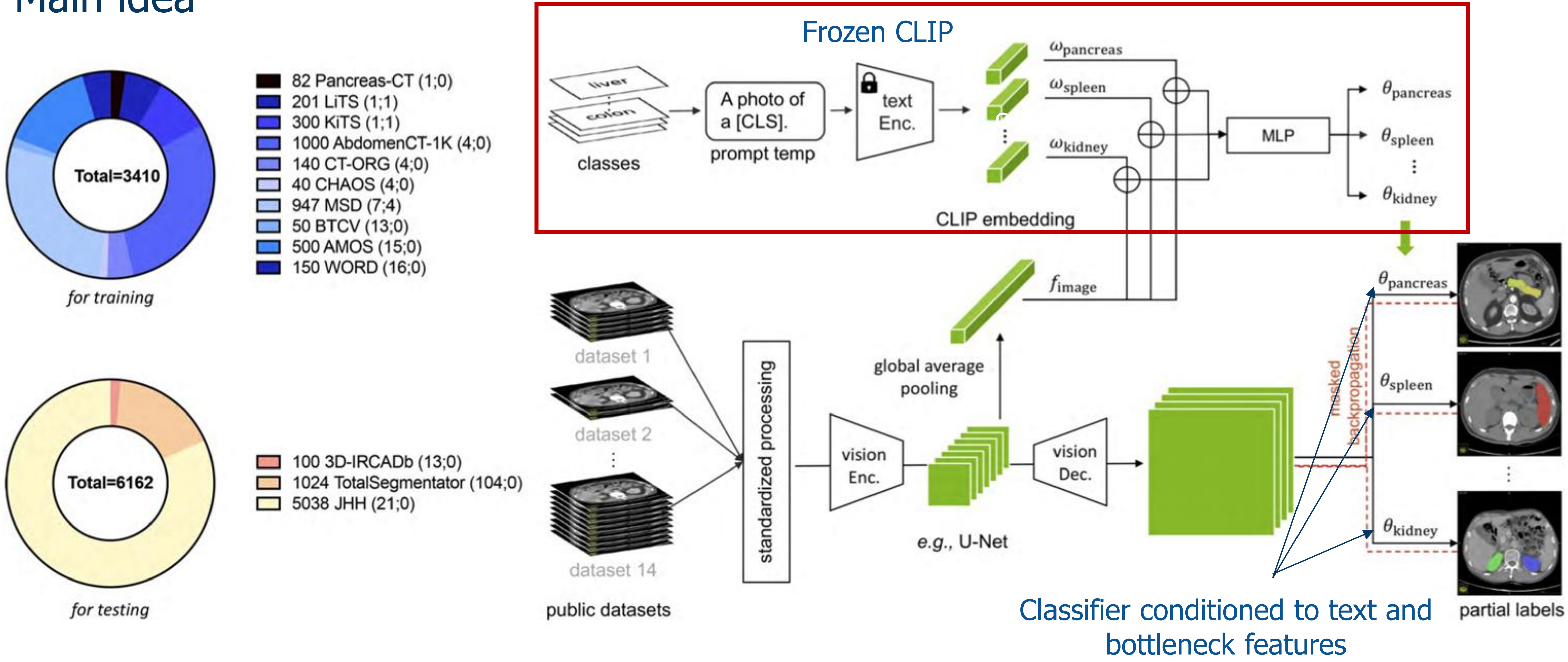
Zero-shot /Adaptation Oriented (3D Data)

CLIP-Driven

SuPreM

How to pre-train? CLIP-Driven

Main idea



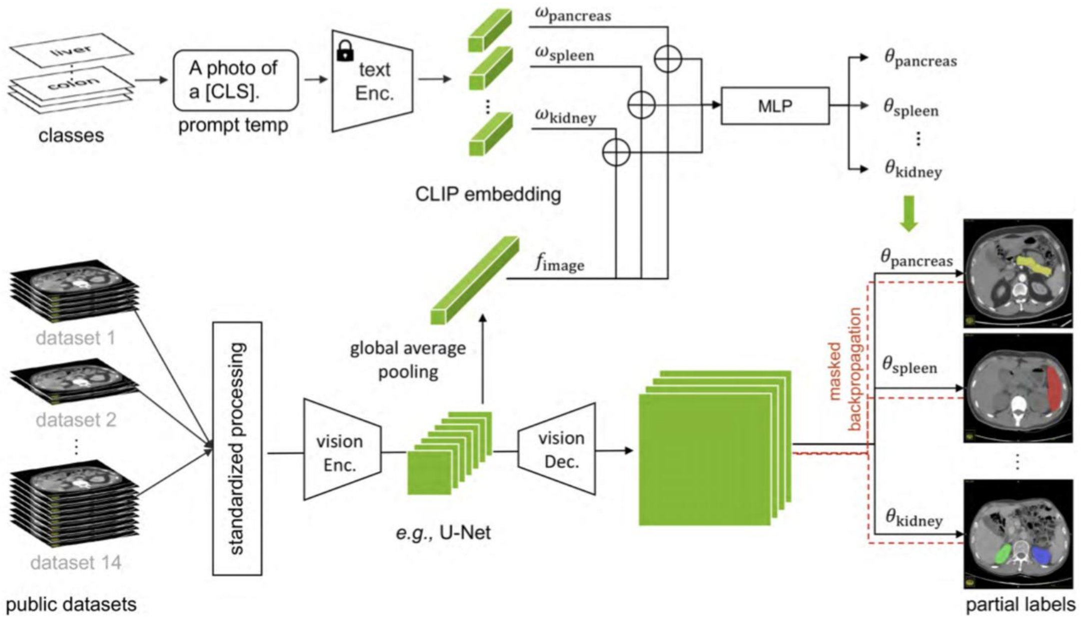
Liu *et al*. CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection. ICCV'23.

Zero-shot /Adaptation Oriented (3D Data)

CLIP-Driven

SuPreM

How to pre-train? CLIP-Driven



Liu et *al*. CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection. ICCV'23.

Zero-shot /Adaptation Oriented (3D Data)

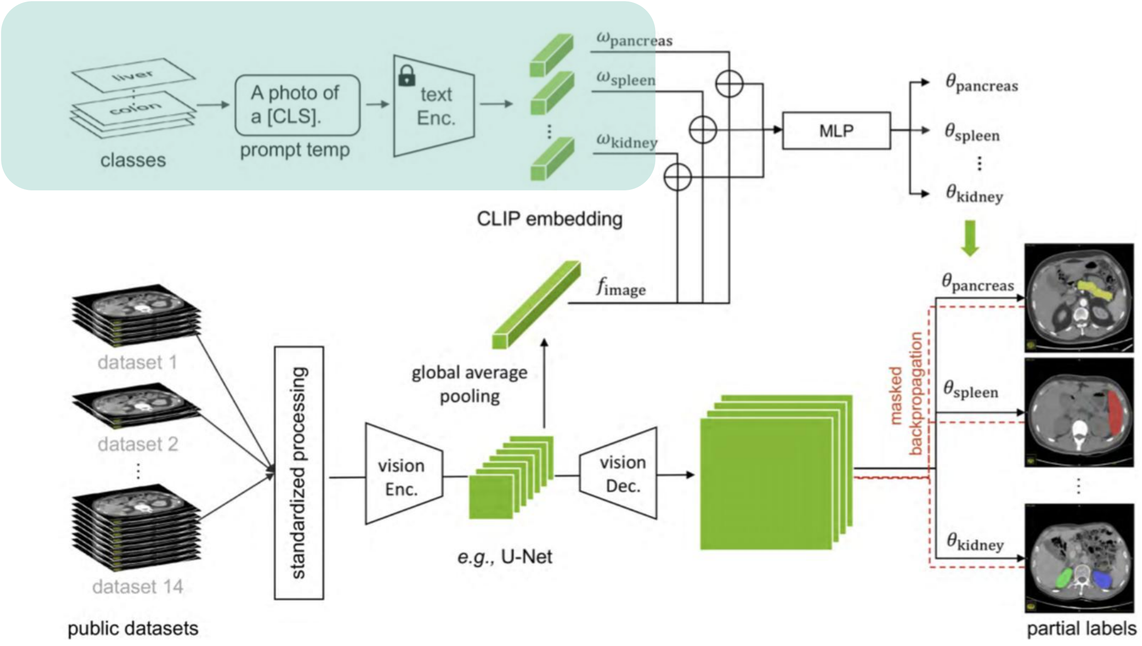
CLIP-Driven

SuPreM

How to pre-train? CLIP-Driven

Text branch
(generates text embedding for class k)

$$\mathbf{w}_k$$



Liu et *al*. CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection. ICCV'23.

Zero-shot /Adaptation Oriented (3D Data)

CLIP-Driven

SuPreM

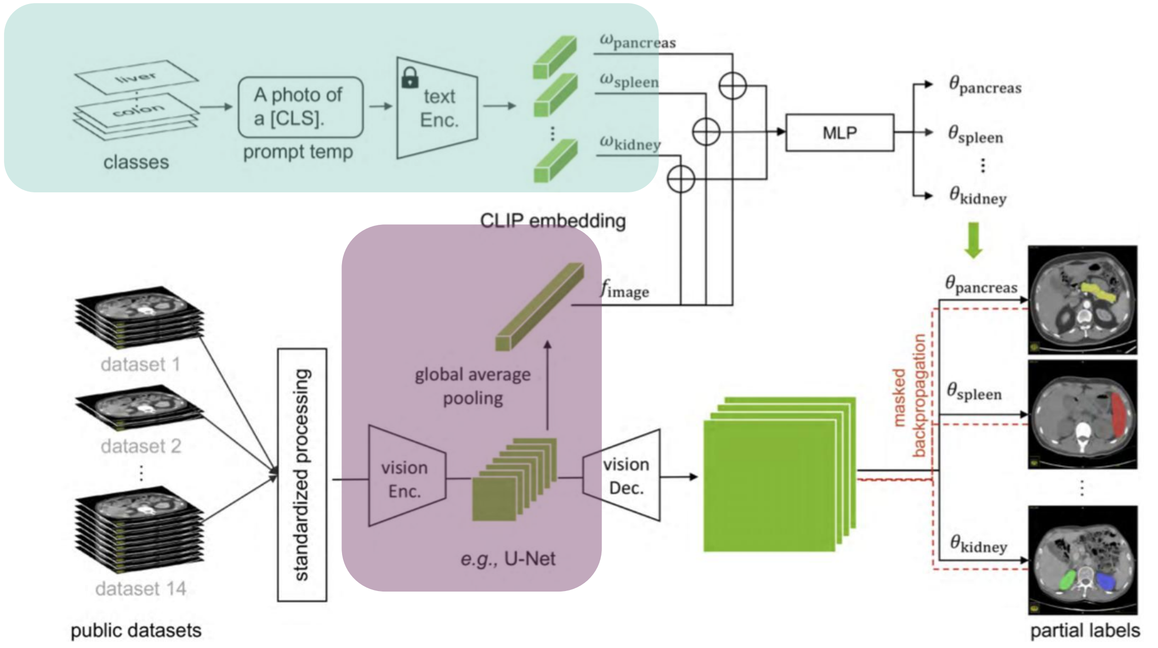
How to pre-train? CLIP-Driven

Text branch
(generates text embedding for class k)

$\mathbf{w}_k$

Visual branch-encoder
(generates visual embedding for volume x)

$\mathbf{f}$



Liu et *al*. CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection. ICCV'23.

Zero-shot /Adaptation Oriented (3D Data)

CLIP-Driven

SuPreM

How to pre-train? CLIP-Driven

Text branch
(generates text embedding for class k)

$$\mathbf{w}_k$$

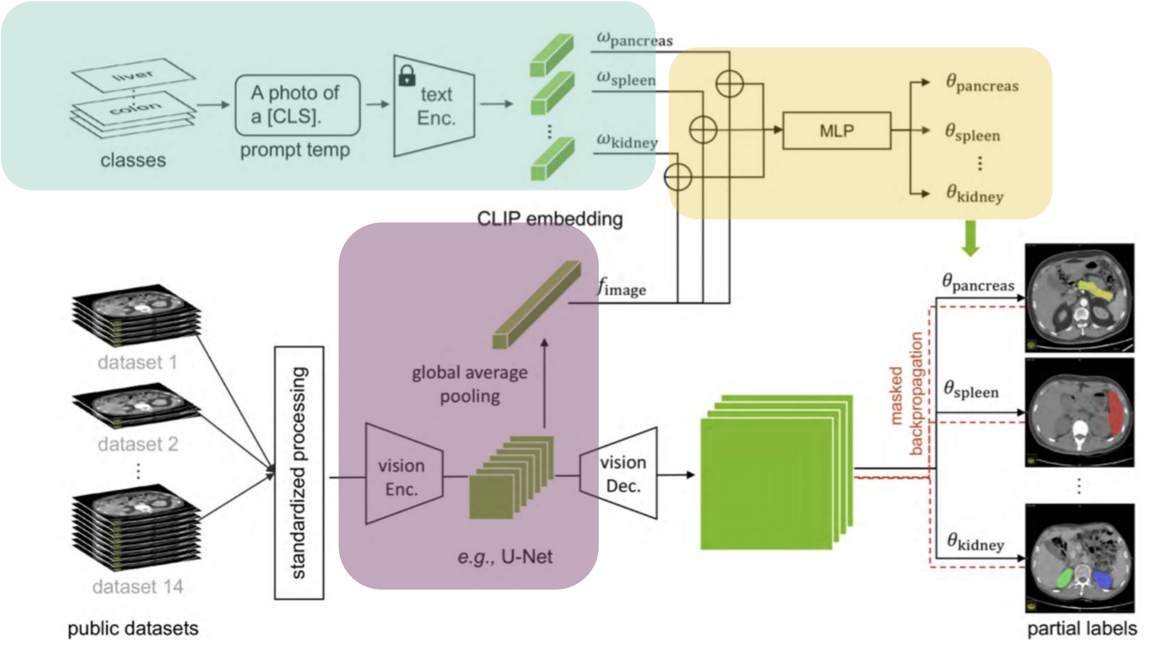
Visual branch-encoder
(generates visual embedding for volume x)

$$\mathbf{f}$$

Text-based controller MLP
(generates class parameters)

$$\boldsymbol{\theta}_{\boldsymbol{k}} = MLP(\mathbf{w}_k \oplus \mathbf{f})$$

$$\boldsymbol{\theta}_k = \{\boldsymbol{\theta}_{k_1}, \boldsymbol{\theta}_{k_2}, \boldsymbol{\theta}_{k_3}\}$$



Liu et *al*. CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection. ICCV’23.

Zero-shot /Adaptation Oriented (3D Data)

CLIP-Driven

SuPreM

How to pre-train? CLIP-Driven

Text branch
(generates text embedding for class k)

$$\mathbf{w}_k$$

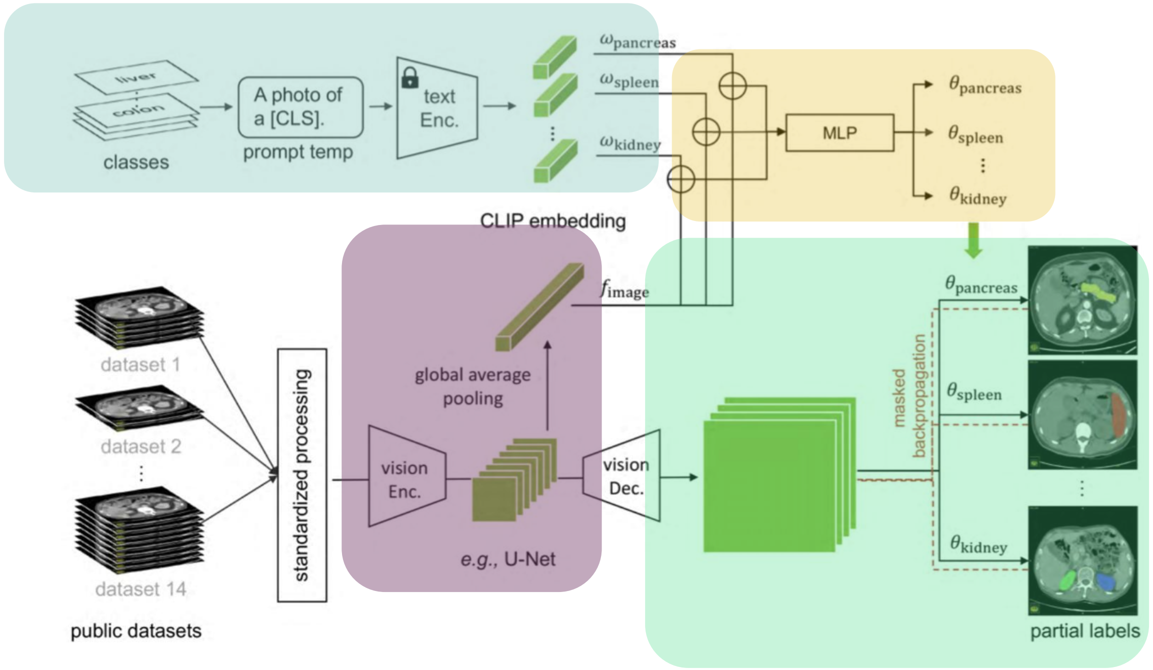
Visual branch-encoder
(generates visual embedding for volume x)

$$\mathbf{f}$$

Text-based controller MLP
(generates class parameters)

$$\boldsymbol{\theta}_{\boldsymbol{k}} = MLP(\mathbf{w}_k \oplus \mathbf{f})$$

$$\boldsymbol{\theta}_k = \{\boldsymbol{\theta}_{k_1}, \boldsymbol{\theta}_{k_2}, \boldsymbol{\theta}_{k_3}\}$$



Visual branch-decoder
(generates visual embedding for image x)

$$\mathbf{P}_k = \mathrm{sigmoid}(((\mathbf{F} * \boldsymbol{\theta}_{k_1}) * \boldsymbol{\theta}_{k_2}) * \boldsymbol{\theta}_{k_3})$$

Liu et *al*. CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection. ICCV'23.

Zero-shot /Adaptation Oriented (3D Data)

CLIP-Driven

SuPreM

How to pre-train? CLIP-Driven

Text branch
(generates text embedding for class k)

$$\mathbf{w}_k$$

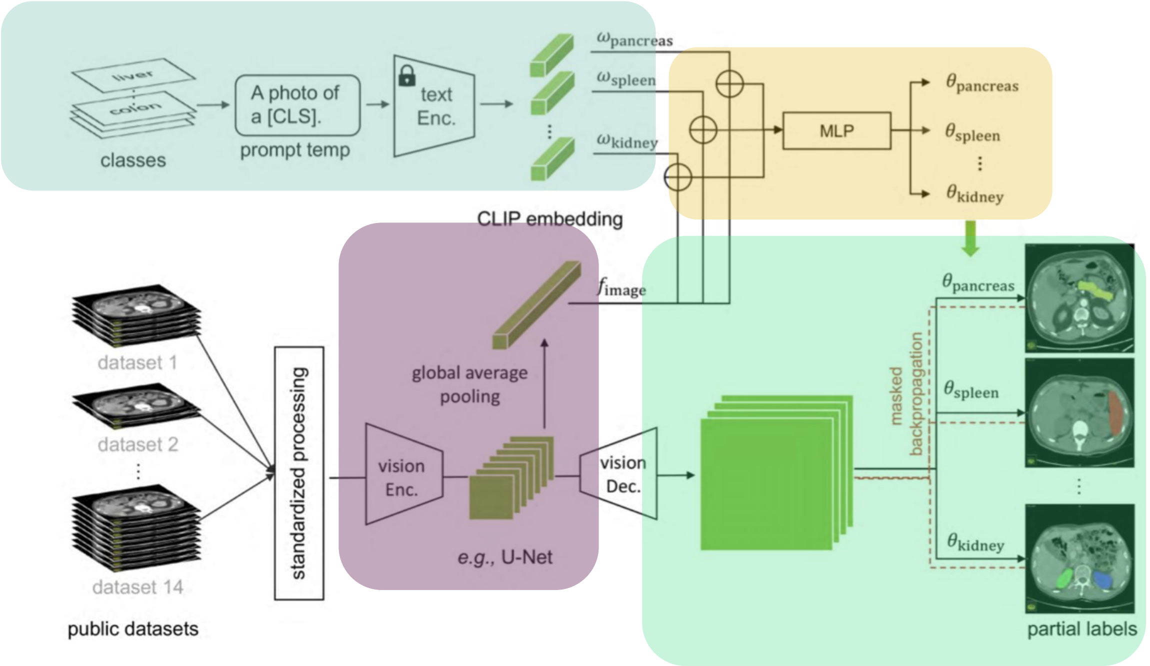
Visual branch-encoder
(generates visual embedding for volume x)

$$\mathbf{f}$$

Text-based controller MLP
(generates class parameters)

$$\boldsymbol{\theta_k} = MLP(\mathbf{w}_k \oplus \mathbf{f})$$

$$\boldsymbol{\theta}_k = \{\boldsymbol{\theta}_{k_1}, \boldsymbol{\theta}_{k_2}, \boldsymbol{\theta}_{k_3}\}$$



Visual branch-decoder
(generates visual embedding for image x)

$$\mathbf{P}_k = \text{sigmoid}(((\mathbf{F} * \boldsymbol{\theta}_{k_1}) * \boldsymbol{\theta}_{k_2}) * \boldsymbol{\theta}_{k_3})$$

Training loss

$$\mathcal{L} = \sum_{k=1}^{K} \mathbf{1}_{\{k \in y\}} \cdot \text{BCE}_k$$

Liu et *al*. CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection. ICCV'23.

Zero-shot /Adaptation Oriented (3D Data)

CLIP-Driven

SuPreM

How to pre-train? CLIP-Driven

Text branch
(generates text embedding for class k)

$$\mathbf{w}_k$$

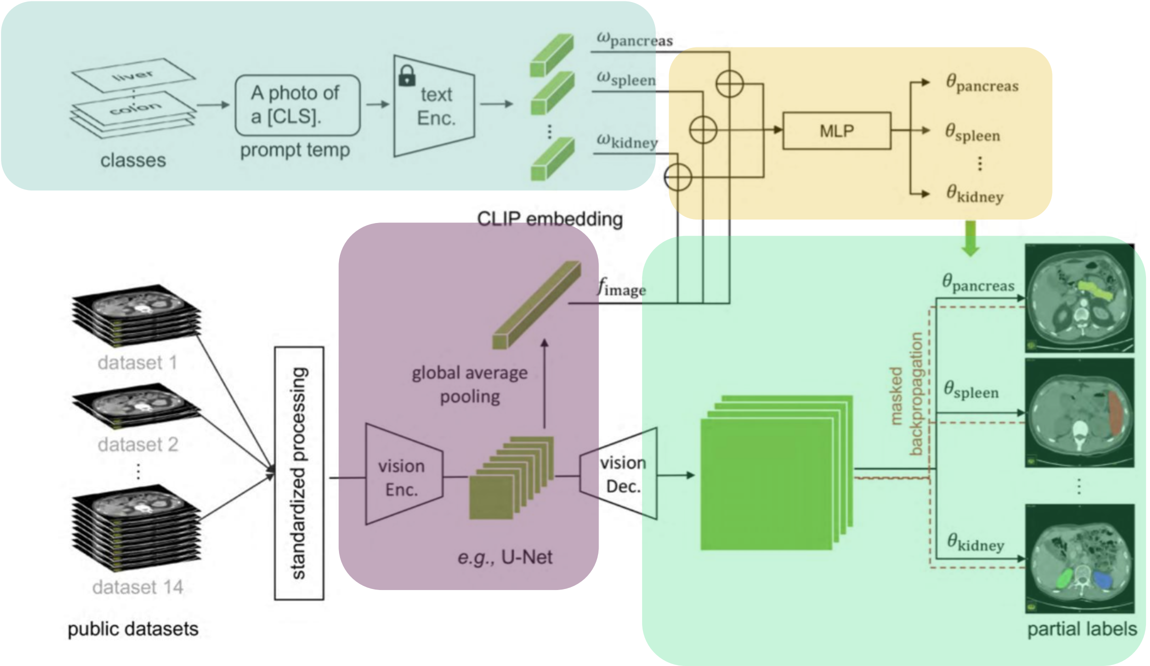
Visual branch-encoder
(generates visual embedding for volume x)

$$\mathbf{f}$$

Text-based controller MLP
(generates class parameters)

$$\boldsymbol{\theta_k} = MLP(\mathbf{w}_k \oplus \mathbf{f})$$

$$\boldsymbol{\theta}_k = \{\boldsymbol{\theta}_{k_1}, \boldsymbol{\theta}_{k_2}, \boldsymbol{\theta}_{k_3}\}$$



Visual branch-decoder
(generates visual embedding for image x)

$$\mathbf{P}_k = \mathrm{sigmoid}(((\mathbf{F} * \boldsymbol{\theta}_{k_1}) * \boldsymbol{\theta}_{k_2}) * \boldsymbol{\theta}_{k_3})$$

Training loss

$$\mathcal{L} = \sum_{k=1}^{K} \mathbf{1}_{\{k \in y\}} \cdot \mathrm{BCE}_k$$

→ **How can the text part contribute if using a generalist model?**

Liu et *al*. CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection. ICCV'23.

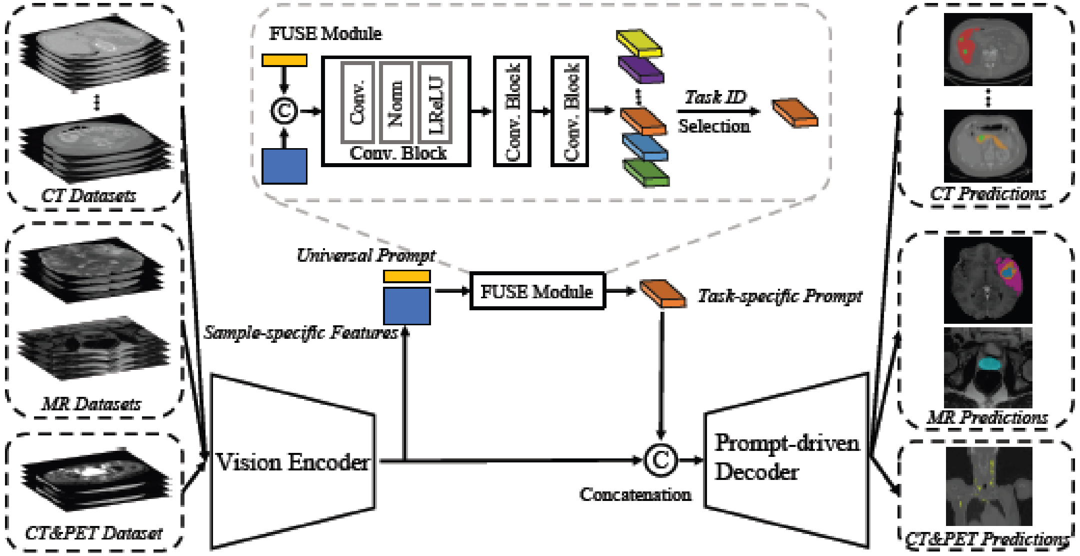
Zero-shot /Adaptation Oriented (3D Data)

UniSeg

Hermes

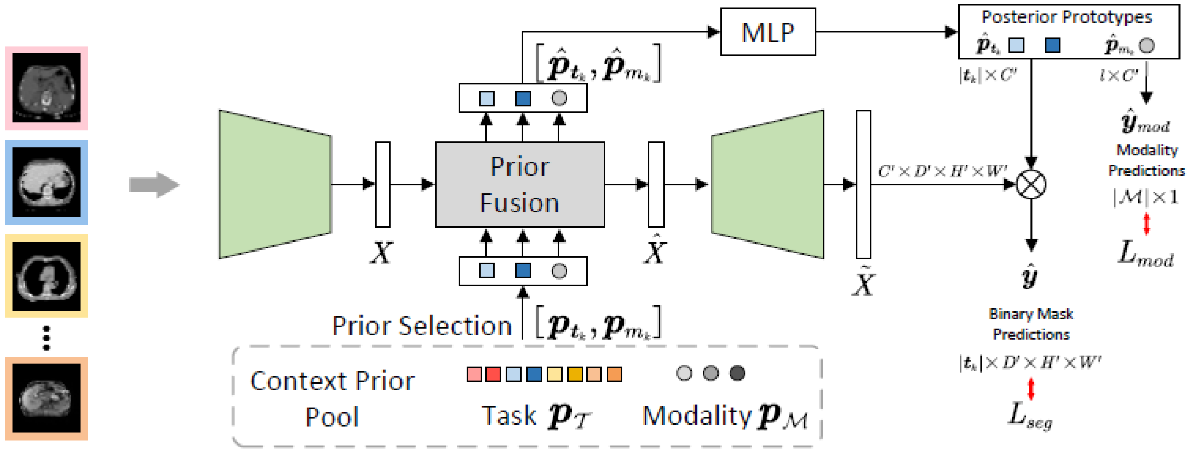
How to pre-train? Prompt-Driven

Main idea



Ye et *al*. UniSeg: A Prompt-driven Universal Segmentation Model as well as A Strong Representation Learner. MICCAI'23.

- **Objective**: condition the segmentation to high level features related to **tasks/domains.**
- **Prompt selection** is a learnable operations to operate during **inference**.



Gao et *al*. Training Like a Medical Resident: Context-Prior Learning Toward Universal Medical Image Segmentation. CVPR'24.

Zero-shot /Adaptation Oriented (3D Data)

UniSeg

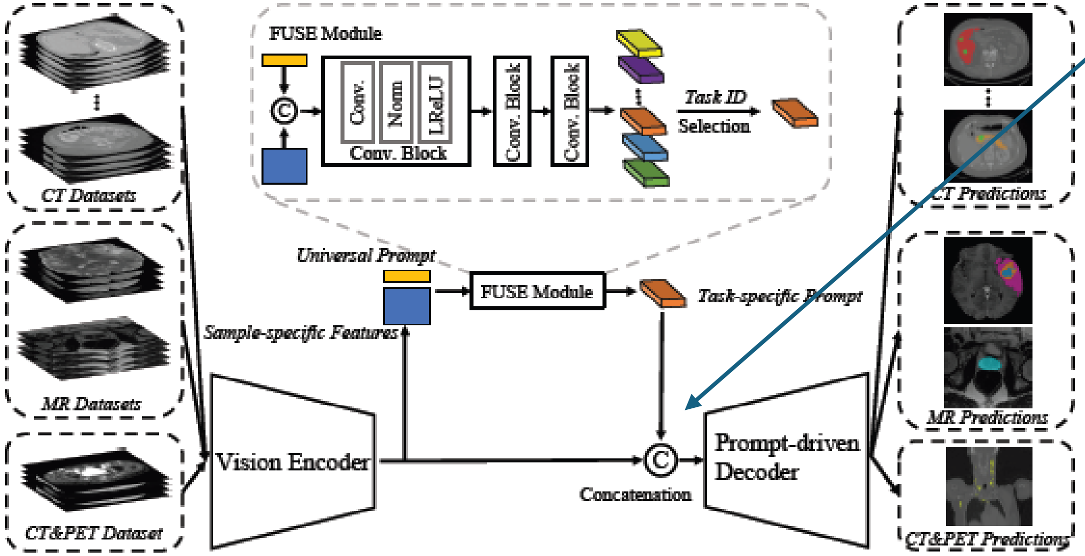
Hermes

How to pre-train? Prompt-Driven

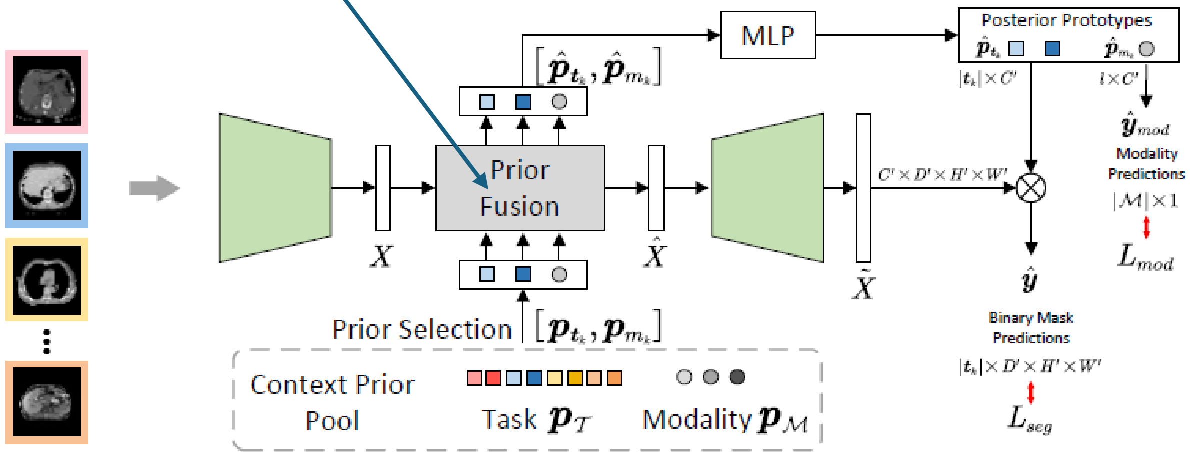
Main idea

Conditioning on decoder path

Conditioning on classifier



Ye et *al*. UniSeg: A Prompt-driven Universal Segmentation Model as well as A Strong Representation Learner. MICCAI'23.



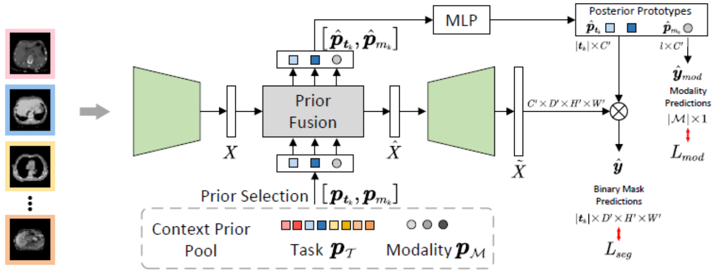
Gao et *al*. Training Like a Medical Resident: Context-Prior Learning Toward Universal Medical Image Segmentation. CVPR'24.

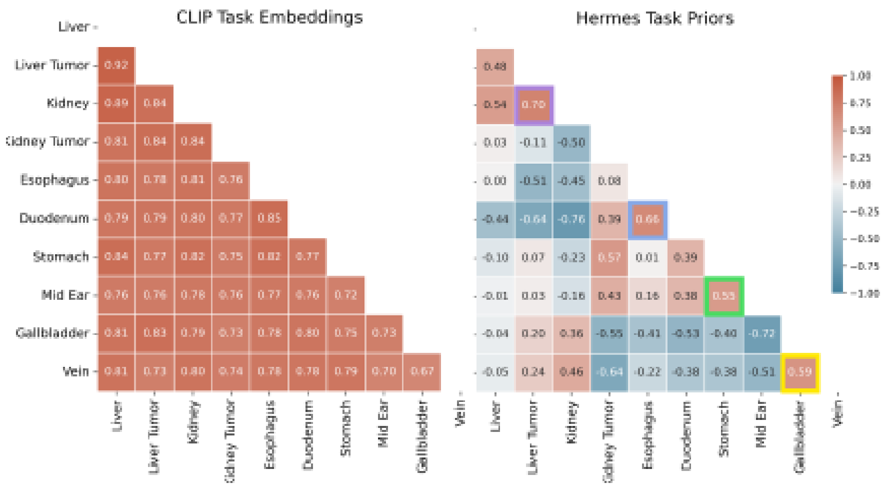
Zero-shot /Adaptation Oriented (3D Data)

UniSeg

Hermes

How to pre-train? Prompt-Driven





Prompt Similarity among tasks

Setting	Model	1%		10%		50%		100%	
		Pan	Tumor	Pan	Tumor	Pan	Tumor	Pan	Tumor
Scratch	ResUNet	44.60	7.67	74.47	23.90	78.89	44.52	80.45	51.06
Transfer	ResUNet (AMOS CT)	56.08	8.31	77.15	25.53	80.53	46.16	81.23	52.21
	ResUNet (KiTS)	52.68	9.28	75.11	27.33	79.07	45.72	79.23	50.64
	DeSD [60] (10,594 CT)	67.82	13.89	78.11	35.82	80.95	50.23	81.97	59.11
	DoDNet [63]	66.62	11.97	76.83	31.92	80.82	47.79	81.41	53.62
	CLIP-Driven [44]	67.95	12.12	77.49	32.37	80.92	48.92	81.45	54.71
	UniSeg [61]	69.05	12.35	77.33	33.87	80.93	49.63	81.96	55.58
	Hermes-R	**72.71**	**16.73**	**79.12**	**44.31**	**81.14**	**55.31**	**82.73**	**61.41**

Gao et *al*. Training Like a Medical Resident: Context-Prior Learning Toward Universal Medical Image Segmentation. CVPR'24.

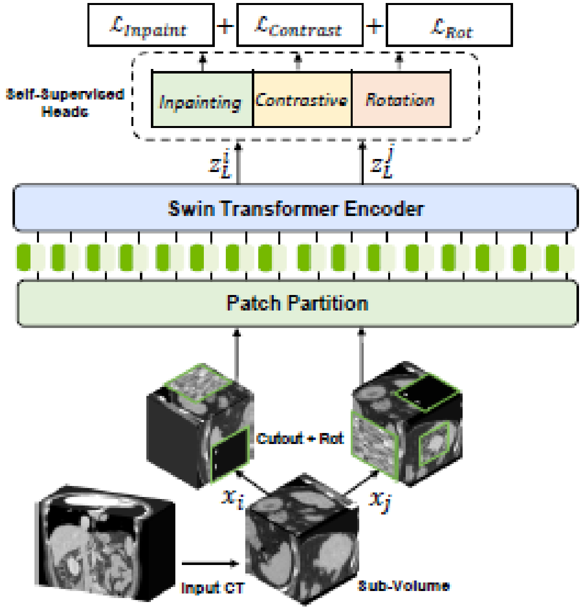
Zero-shot /Adaptation Oriented (3D Data)

How to pretrain? Self-supervised pre-training

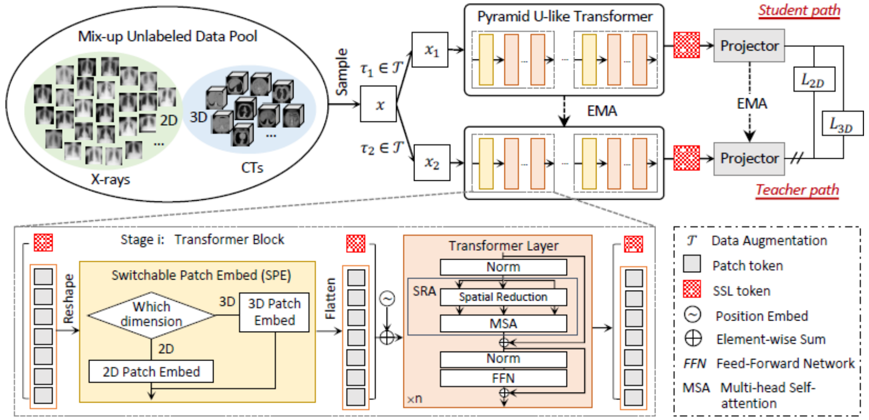
→ Producing quality annotations in volumetric scans is expensive and laborious.

→ Large amounts of unlabeled data are available.
(current self-supervised models are pre-trained with more than 5000 CT scans)

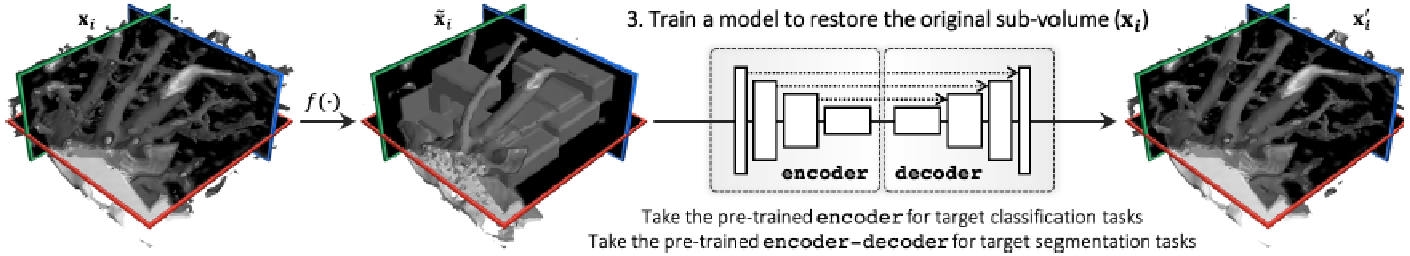
→ Different pretext tasks.



Tang et *al*. Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis. CVPR'22.



Xie et *al*. UniMiSS: Universal Medical Self-Supervised Learning via Breaking Dimensionality Barrier. ECCV'22.



Zhou et *al*. Model Genesis. MedIA'21.

Zero-shot /Adaptation Oriented (3D Data)

SuPreM

MultiTalent

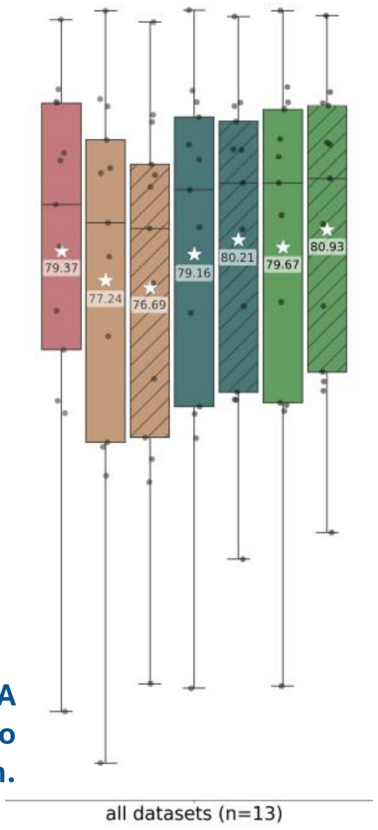
Benefits of foundation models?

→ Transferability via **full fine-tuning** of the pre-trained model.
→ Access to **hundreds of labeled volumes for adaptation**.

	name	backbone	params	pre-trained data	performance†
self-supervised	Models Genesis (Zhou et al., 2019)	U-Net	19.08M	623 CT volumes	90.1
	UniMiSS (Xie et al., 2022)	nnU-Net	61.79M	5,022 CT&MRI volumes	92.9
	NV*	Swin UNETR	62.19M	1,000 CT volumes	93.2
	NV*	Swin UNETR	62.19M	3,000 CT volumes	93.4
	NV (Tang et al., 2022)	Swin UNETR	62.19M	5,050 CT volumes	93.8
	NV*	Swin UNETR	62.19M	5,050 CT volumes	94.2
	NV*	Swin UNETR	62.19M	9,262 CT volumes	94.3
supervised	Med3D (Chen et al., 2019b)	Residual U-Net	85.75M	1,638 CT volumes	91.4
	DoDNet (Zhang et al., 2021)	U-Net	17.29M	920 CT volumes	93.8
	DoDNet*	U-Net	17.29M	920 CT volumes	94.4
	Universal Model (Liu et al., 2023b)	U-Net	19.08M	2,100 CT volumes	-
	Universal Model (Liu et al., 2023b)	Swin UNETR	62.19M	2,100 CT volumes	94.1
	SuPreM*	U-Net	19.08M	2,100 CT volumes	**95.4**
	SuPreM*	Swin UNETR	62.19M	2,100 CT volumes	94.6
	SuPreM*	SegResNet	470.13M	2,100 CT volumes	94.0

+0.8%

Li et *al*. How Well Do Supervised 3D Models Transfer to Medical Imaging Tasks?. ICLR'24.



+1.0%

Ulrich et *al*. MultiTalent: A Multi-Dataset Approach to Medical Image Segmentation. MICCAI'23.

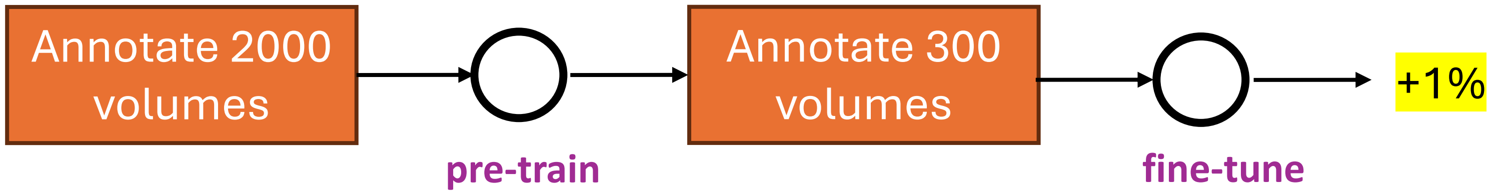
Zero-shot /Adaptation Oriented (3D Data)

SuPreM

MultiTalent

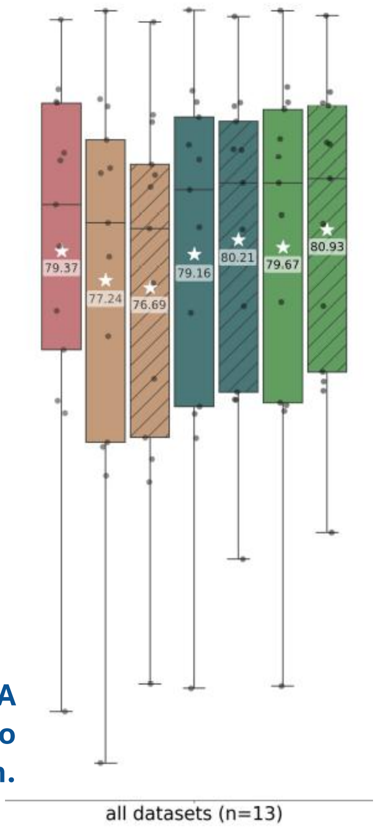
Benefits of foundation models?

→ Transferability via **full fine-tuning** of the pre-trained model.
→ Access to **hundreds of labeled volumes for adaptation**.



	name	backbone	params	pre-trained data	performance†
self-supervised	Models Genesis (Zhou et al., 2019)	U-Net	19.08M	623 CT volumes	90.1
	UniMiSS (Xie et al., 2022)	nnU-Net	61.79M	5,022 CT&MRI volumes	92.9
	NV*	Swin UNETR	62.19M	1,000 CT volumes	93.2
	NV*	Swin UNETR	62.19M	3,000 CT volumes	93.4
	NV (Tang et al., 2022)	Swin UNETR	62.19M	5,050 CT volumes	93.8
	NV*	Swin UNETR	62.19M	5,050 CT volumes	94.2
	NV*	Swin UNETR	62.19M	9,262 CT volumes	94.3
supervised	Med3D (Chen et al., 2019b)	Residual U-Net	85.75M	1,638 CT volumes	91.4
	DoDNet (Zhang et al., 2021)	U-Net	17.29M	920 CT volumes	93.8
	DoDNet*	U-Net	17.29M	920 CT volumes	94.4
	Universal Model (Liu et al., 2023b)	U-Net	19.08M	2,100 CT volumes	-
	Universal Model (Liu et al., 2023b)	Swin UNETR	62.19M	2,100 CT volumes	94.1
	SuPreM*	U-Net	19.08M	2,100 CT volumes	**95.4**
	SuPreM*	Swin UNETR	62.19M	2,100 CT volumes	94.6
	SuPreM*	SegResNet	470.13M	2,100 CT volumes	94.0

Li et *al*. How Well Do Supervised 3D Models Transfer to Medical Imaging Tasks?. ICLR'24.



Ulrich et *al*. MultiTalent: A Multi-Dataset Approach to Medical Image Segmentation. MICCAI'23.

Zero-shot /Adaptation Oriented (3D Data)

SuPreM

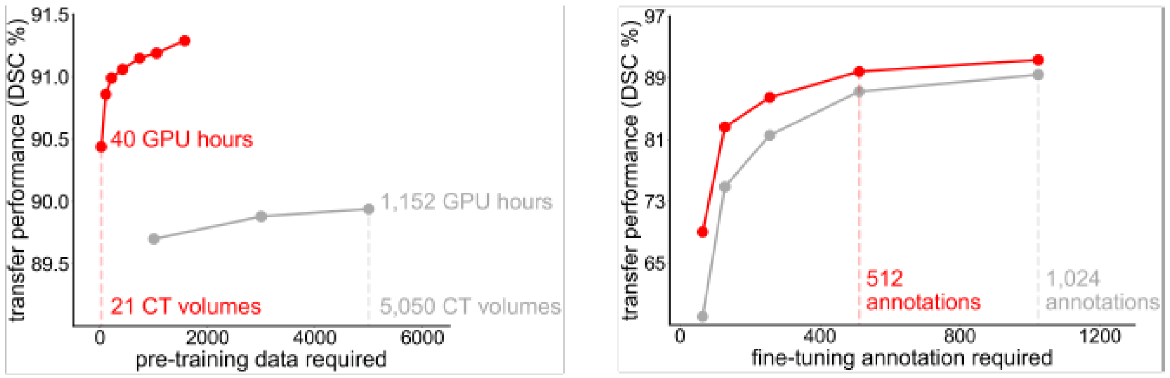
Benefits of foundation models

→ SuPreM models are pre-trained on a curated dataset with 25 fully-annotated structures.

Li et al. AdbomenAtlas: A Large Scale Detailed Annotated and Multi Center Dataset for Efficient Transfer Learning and Open Algorithmic Benchmarking. MedIA'24.

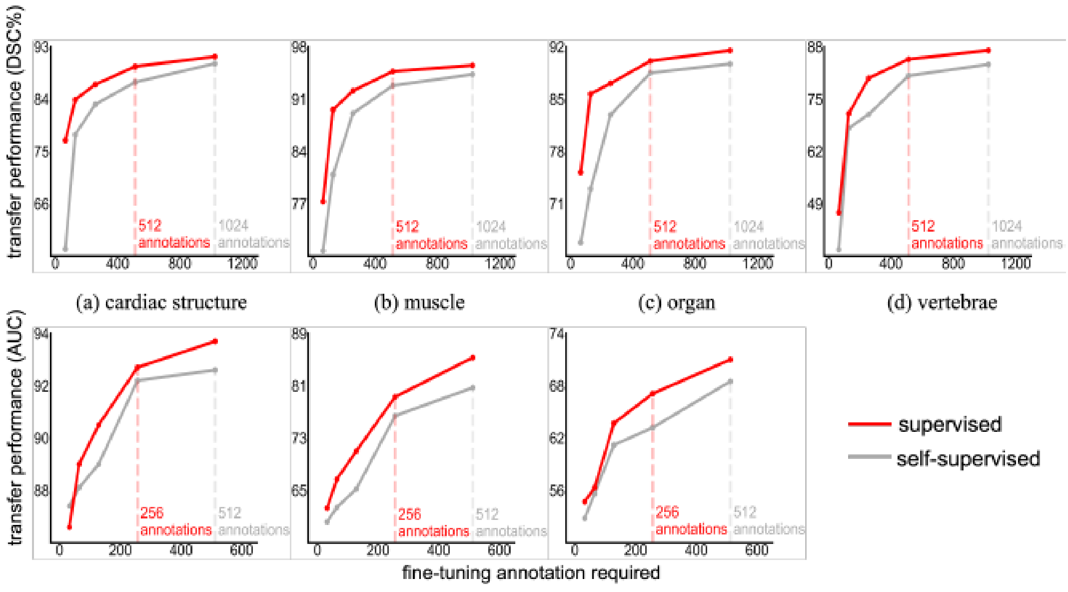
→ **Supervised pre-training is orders of magnitude more data-efficient than self-supervision.**

→ **This holds even when transferring to unseen structures.**



(a) data & computational efficiency in *pre-training*

(b) annotation & learning efficiency in *fine-tuning*



(a) cardiac structure (b) muscle (c) organ (d) vertebrae

(e) PDAC (f) Cyst (g) PanNET

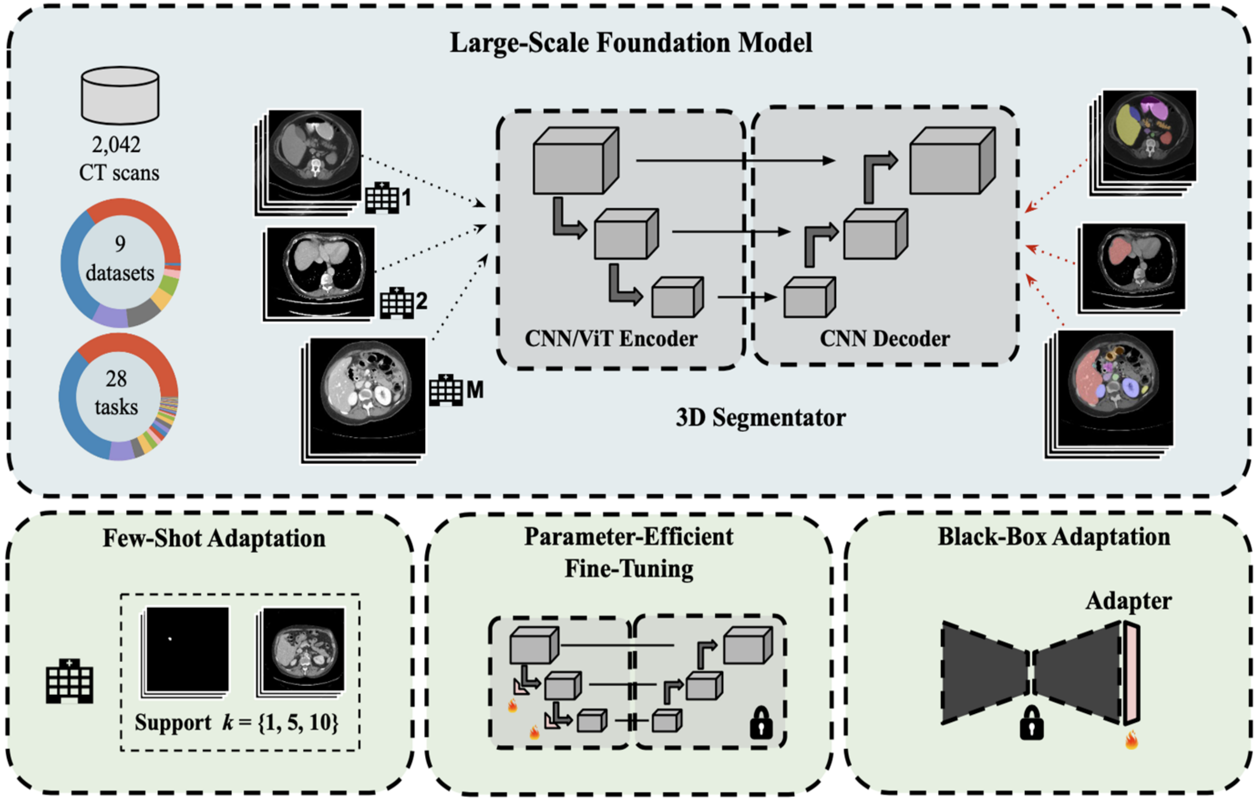
Li et al. How Well Do Supervised 3D Models Transfer to Medical Imaging Tasks?. ICLR'24.

Zero-shot /Adaptation Oriented (3D Data)

FSEFT

Few-Shot Efficient Fine-Tuning

Main idea (how to adapt a pre-trained large-scale model efficiently)



Silva-Rodríguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MICCAI W-MedAGI'23 (Extension Under Review).

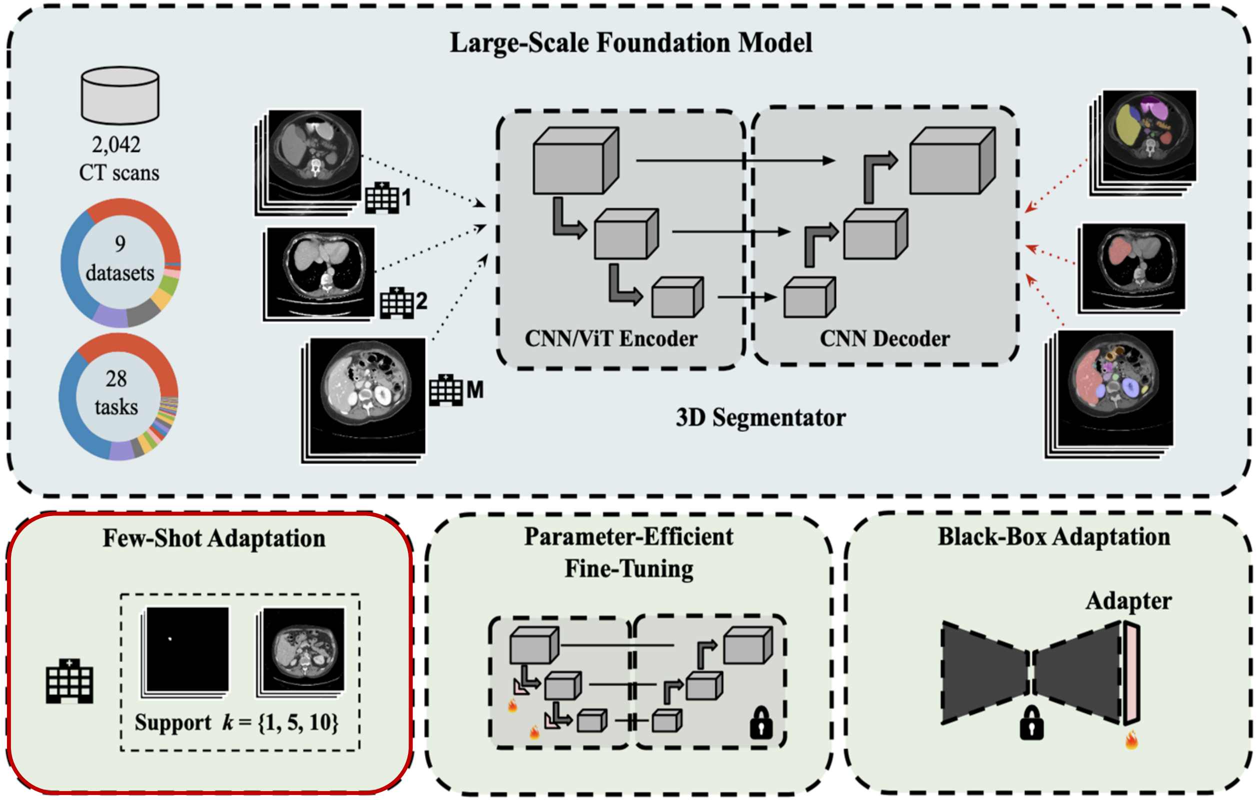
Zero-shot /Adaptation Oriented (3D Data)

FSEFT

Few-Shot Efficient Fine-Tuning

Main idea (how to adapt a pre-trained large-scale model efficiently)

Presence of few annotated volumes for adaptation



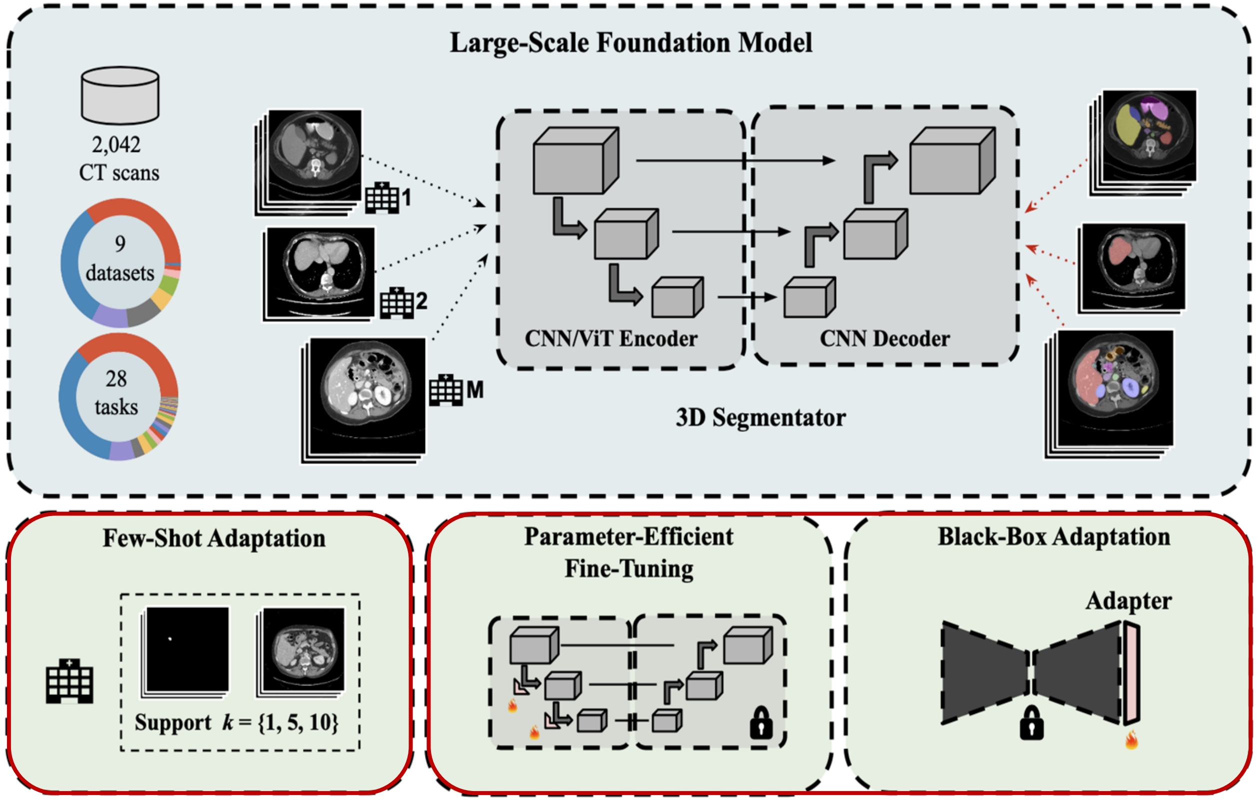
Silva-Rodríguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MICCAI W-MedAGI'23 (Extension Under Review).

Zero-shot /Adaptation Oriented (3D Data)

FSEFT

Few-Shot Efficient Fine-Tuning

Main idea (how to adapt a pre-trained large-scale model efficiently)



Presence of few annotated volumes for adaptation

Being computationally efficient, allowing for commodity GPUs

Silva-Rodríguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MICCAI W-MedAGI'23 (Extension Under Review).

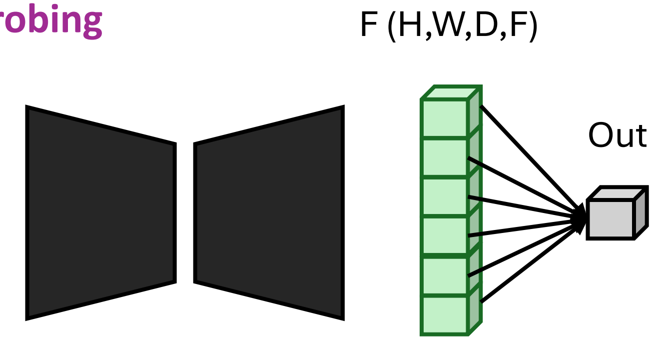
Zero-shot /Adaptation Oriented (3D Data)

FSEFT

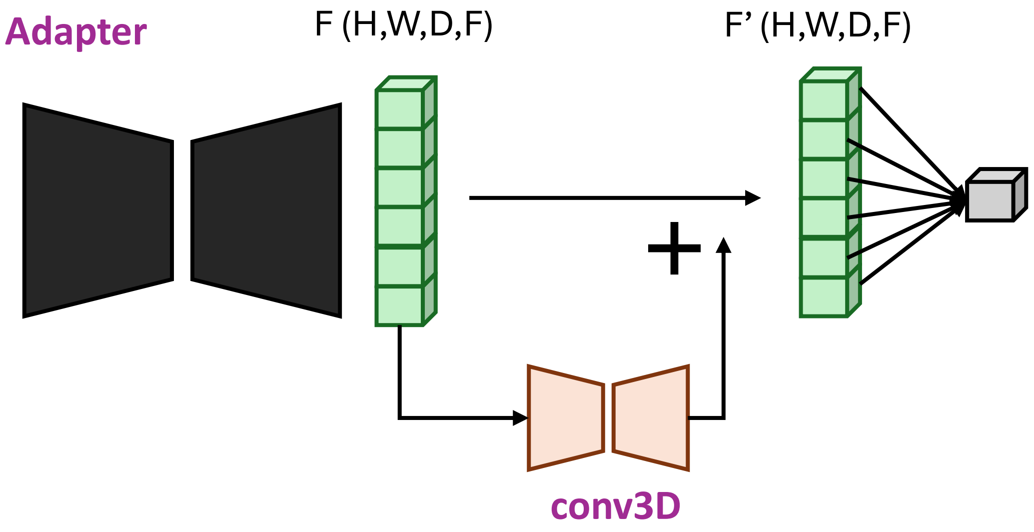
Few-Shot Efficient Fine-Tuning

Black-box Adapters

Linear Probing



Spatial Adapter



Silva-Rodríguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MICCAI W-MedAGI'23 (Extension Under Review).

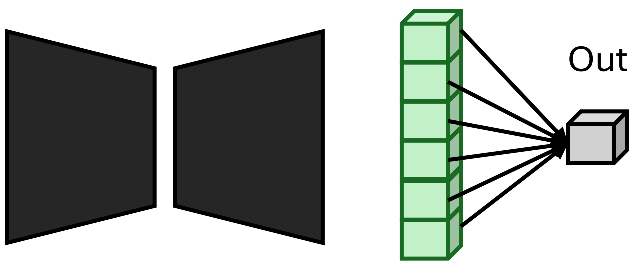
Zero-shot /Adaptation Oriented (3D Data)

FSEFT

Few-Shot Efficient Fine-Tuning

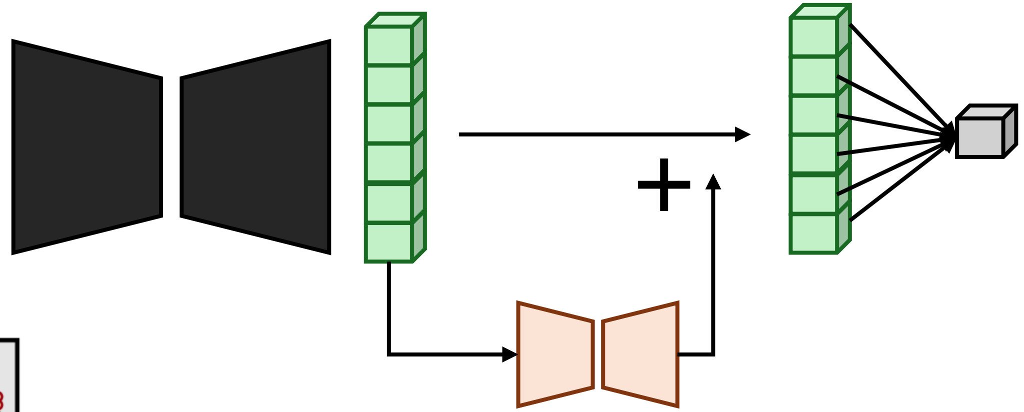
Black-box Adapters

Linear Probing

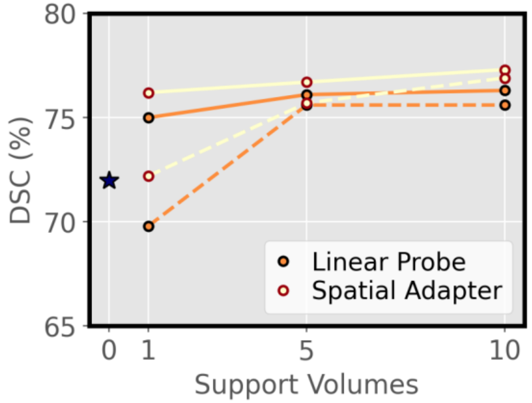


A

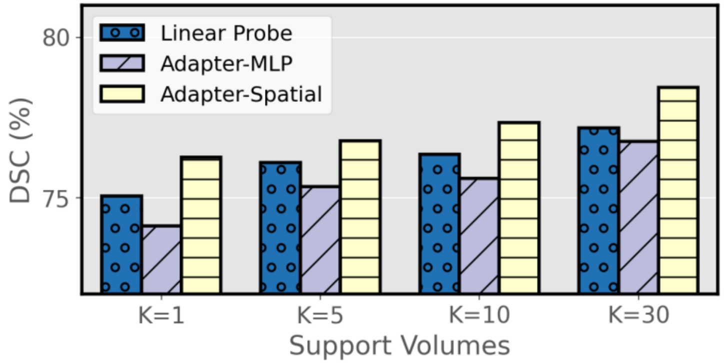
Spatial Adapter



Initialization



MLP vs. Spatial



Silva-Rodríguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MICCAI W-MedAGI’23 (Extension Under Review).

Zero-shot /Adaptation Oriented (3D Data)

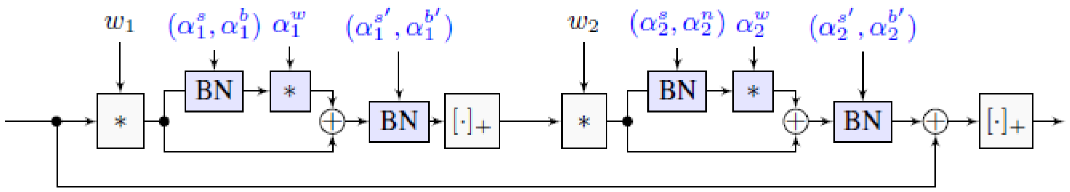
FSEFT

Few-Shot Efficient Fine-Tuning

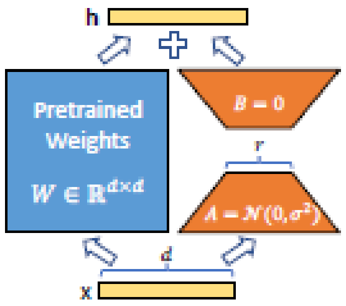
Parameter-Efficient Fine-Tuning

Selective

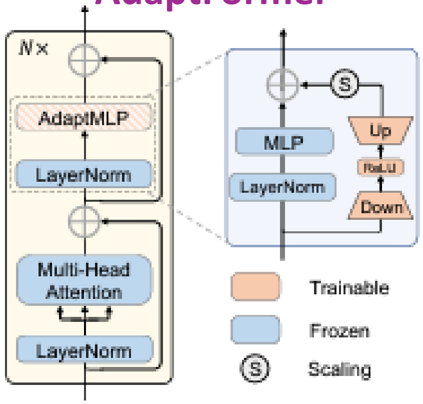
Adapter-CNN



LoRA



AdaptFormer



Silva-Rodríguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MICCAI W-MedAGI'23 (Extension Under Review).

Zero-shot /Adaptation Oriented (3D Data)

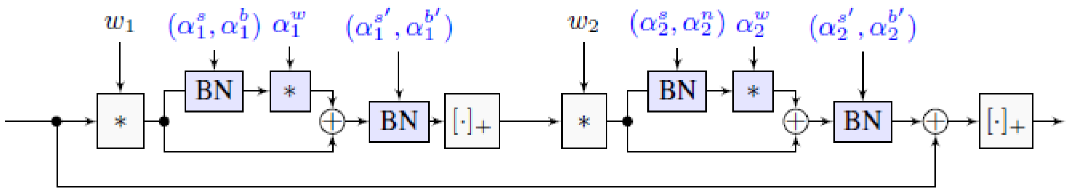
FSEFT

Few-Shot Efficient Fine-Tuning

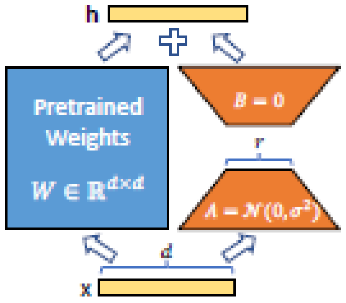
Parameter-Efficient Fine-Tuning

<u>Selective</u>

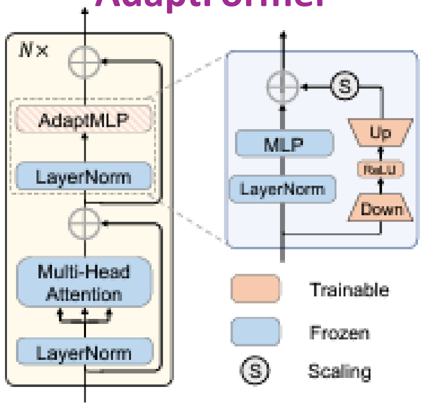
Adapter-CNN



LoRA



AdaptFormer



<u>Additive</u>

Affine-BN/LN

$$y = \frac{x - \mathrm{E}[x]}{\sqrt{\mathrm{Var}[x] + \epsilon}} * \gamma + \beta$$

BitFit (only-bias)

Silva-Rodríguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MICCAI W-MedAGI'23 (Extension Under Review).

Zero-shot / Adaptation Oriented (3D Data)

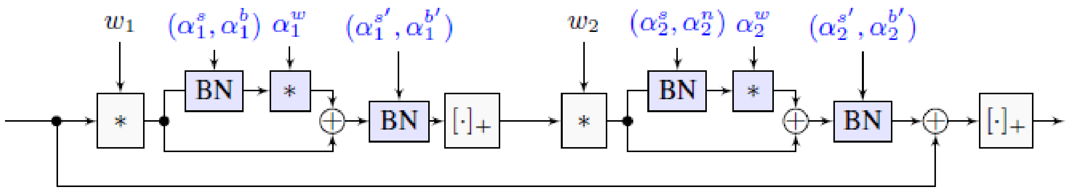
FSEFT

Few-Shot Efficient Fine-Tuning

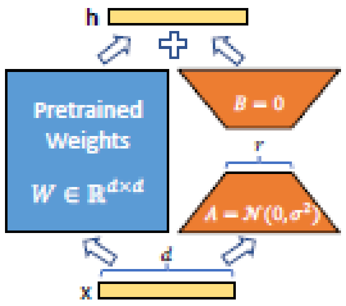
Parameter-Efficient Fine-Tuning

Selective

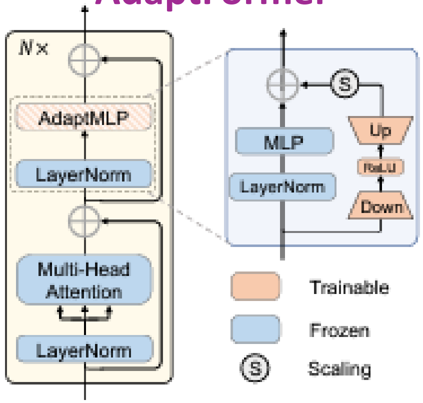
Adapter-CNN



LoRA



AdaptFormer



Additive

Affine-BN/LN

$$y = \frac{x - \mathrm{E}[x]}{\sqrt{\mathrm{Var}[x] + \epsilon}} * \gamma + \beta$$

BitFit (only-bias)

What to do with the Decoder? (millions of paramerers)

→ **Base** categories: **frozen**.

→ **New** categories: **fine-tuned**.

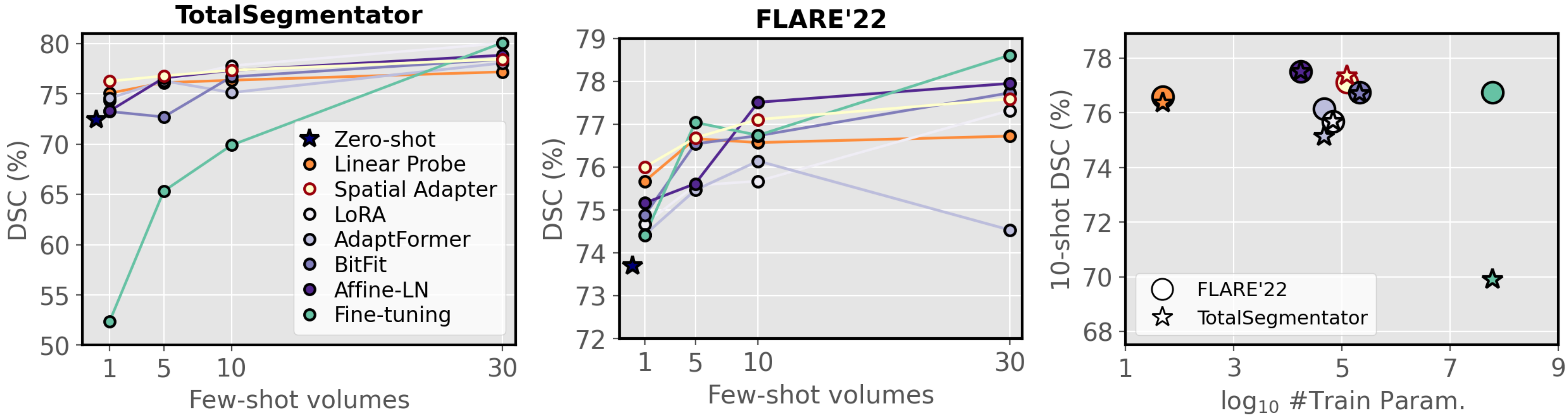
Silva-Rodríguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MICCAI W-MedAGI'23 (Extension Under Review).

Zero-shot /Adaptation Oriented (3D Data)

FSEFT

Few-Shot Efficient Fine-Tuning

Transferability to known tasks (domain shift)



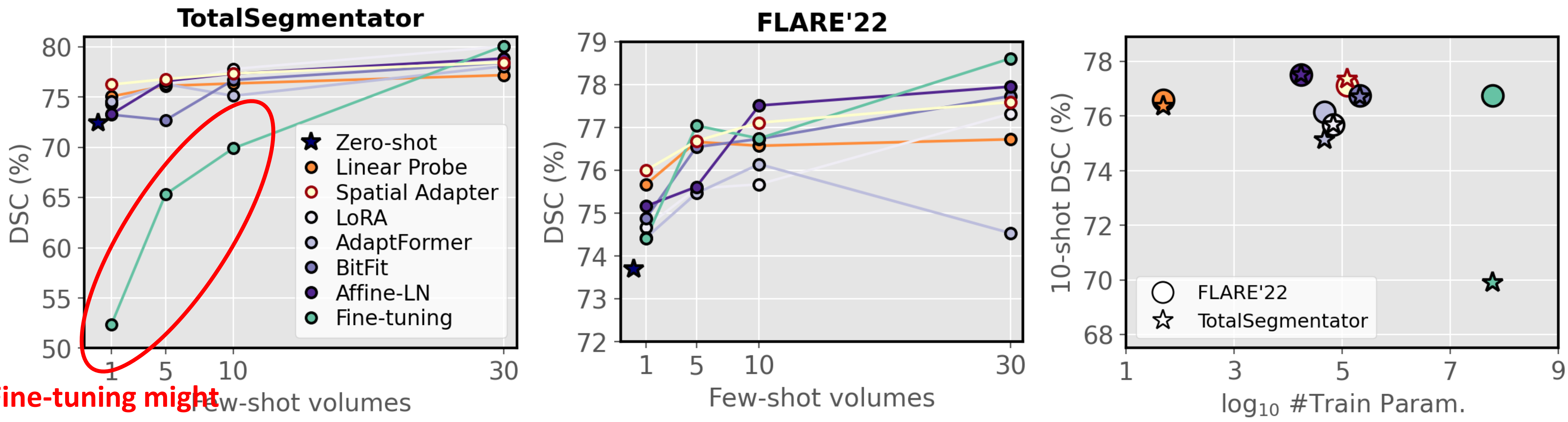
Silva-Rodríguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MICCAI W-MedAGI'23 (Extension Under Review).

Zero-shot /Adaptation Oriented (3D Data)

FSEFT

Few-Shot Efficient Fine-Tuning

Transferability to known tasks (domain shift)



Fine-tuning might distort the model

Silva-Rodríguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MICCAI W-MedAGI'23 (Extension Under Review).

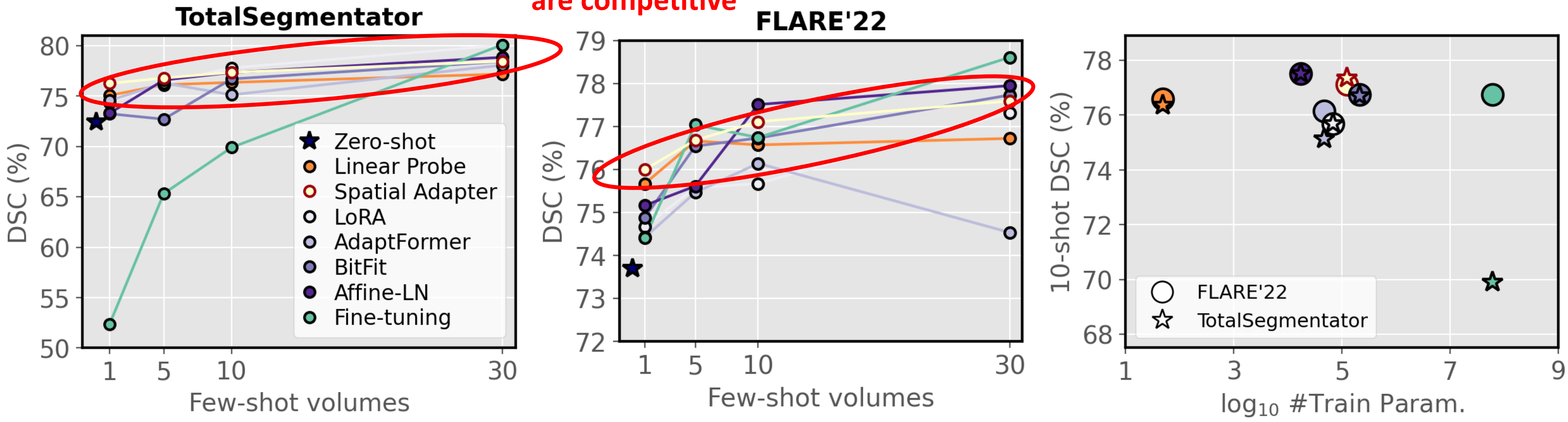
Zero-shot /Adaptation Oriented (3D Data)

FSEFT

Few-Shot Efficient Fine-Tuning

Transferability to known tasks (domain shift)

Black-box methods are competitive



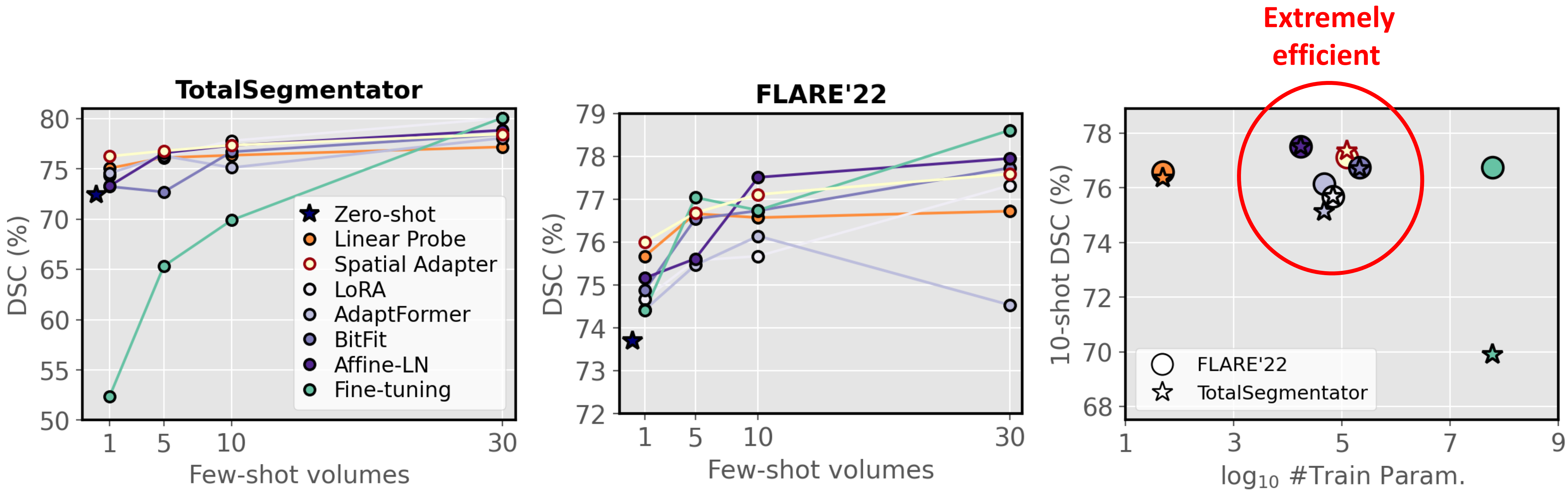
Silva-Rodríguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MICCAI W-MedAGI'23 (Extension Under Review).

Zero-shot / Adaptation Oriented (3D Data)

FSEFT

Few-Shot Efficient Fine-Tuning

Transferability to known tasks (domain shift)



Silva-Rodríguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MICCAI W-MedAGI'23 (Extension Under Review).

Zero-shot /Adaptation Oriented (3D Data)

FSEFT

Few-Shot Efficient Fine-Tuning

Transferability to known tasks (domain shift)

Setting		Method	Spl	lKid	Gall	Eso	Liv	Pan	Sto	Duo	Aor	Avg.
5-shot	PEFT	CNN-Adapter (Rebuffi et al., 2018)	47.69	39.58	40.52	53.05	55.08	43.17	28.47	35.73	84.62	47.55
		Bias (Cai et al., 2020)	71.16	69.54	70.16	55.86	71.03	79.60	51.25	69.04	88.92	69.62
		Affine-BN (Frankle et al., 2021)	69.22	72.33	65.66	52.68	67.61	75.50	45.08	66.52	86.94	66.84
	BB	Linear Probe	93.91	75.59	75.94	50.50	80.29	68.19	57.18	77.18	88.48	74.14
		Spatial Adapter	91.78	77.71	80.89	52.30	90.00	78.83	83.27	80.37	89.08	**80.47**
10-shot	PEFT	CNN-Adapter (Rebuffi et al., 2018)	57.32	61.79	42.96	55.61	52.21	52.77	39.96	34.97	89.26	54.09
		Bias (Cai et al., 2020)	72.79	76.14	83.37	59.65	73.97	79.68	60.65	73.46	92.80	74.72
		Affine-BN (Frankle et al., 2021)	72.15	74.06	77.15	58.65	72.31	77.08	61.74	63.94	92.43	72.17
	BB	Linear Probe	91.22	75.63	77.48	50.02	80.87	69.17	56.28	77.63	85.29	73.73
		Spatial Adapter	95.40	83.76	81.29	52.49	90.75	78.57	81.97	81.09	90.33	**81.74**

(a) 3D-UNet

Setting		Method	Spl	lKid	Gall	Eso	Liv	Pan	Sto	Duo	Aor	Avg.
5-shot	PEFT	BitFit (Ben-Zaken et al., 2021)	88.76	85.91	79.42	50.22	92.17	73.64	62.81	69.30	90.82	77.01
		LoRA (Hu et al., 2022)	61.31	46.52	52.50	46.43	80.50	66.86	38.66	54.15	73.33	57.81
		AdaptFormer (Chen et al., 2022a)	87.57	86.05	60.17	51.79	90.11	76.73	68.29	74.49	93.12	76.48
		Affine-LN (Basu et al., 2024)	88.14	83.81	76.10	50.04	91.89	75.46	64.41	71.91	90.91	76.96
	BB	Linear Probe	94.62	91.86	82.98	49.29	93.54	78.86	72.43	77.30	88.77	81.07
		Spatial Adapter	95.34	88.13	85.08	55.56	94.27	78.84	75.33	78.17	87.40	**82.01**
10-shot	PEFT	BitFit (Ben-Zaken et al., 2021)	95.16	86.54	84.86	56.93	93.58	72.03	69.26	75.47	90.44	80.47
		LoRA (Hu et al., 2022)	63.97	54.53	59.25	55.33	84.03	77.72	58.72	73.89	80.59	67.56
		AdaptFormer (Chen et al., 2022a)	91.36	84.03	77.78	54.10	93.14	76.05	70.08	77.58	93.25	79.71
		Affine-LN (Basu et al., 2024)	87.21	87.36	80.84	55.80	93.65	76.98	66.78	75.66	92.50	79.64
	BB	Linear Probe	95.26	91.63	82.15	52.69	93.37	69.93	71.70	77.20	88.70	80.29
		Spatial Adapter	95.83	89.44	81.61	56.24	94.40	77.69	76.03	79.54	84.66	**81.72**

(b) Swin-UNETR

Black-box methods hold their performance when directly applied to SuPreM models

Silva-Rodríguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MICCAI W-MedAGI'23 (Extension Under Review).

Zero-shot /Adaptation Oriented (3D Data)

FSEFT

Few-Shot Efficient Fine-Tuning

Transferability to novel tasks (new organs)

Setting	Method	Lung*	Heart†	Gluteus‡	Avg.
FULL	Fine-tuning (Tang et al., 2022)	19.59	53.14	55.37	42.70
	Fine-tuning (*Ours*)	**31.01**	60.79	65.35	52.38
PEFT	BitFit (Ben-Zaken et al., 2021)	14.79	48.90	39.43	34.28
	LoRA (Hu et al., 2022)	13.80	50.55	46.36	38.49
	AdaptFormer (Chen et al., 2022a)	18.82	53.35	48.61	40.26
	Affine-LN (Basu et al., 2024)	16.92	58.38	46.07	40.46
	Decoder fine-tuning	25.98	65.69	64.23	51.97
	+BitFit (Ben-Zaken et al., 2021)	26.17	65.78	64.34	52.10
	+LoRA (Hu et al., 2022)	26.16	**76.12**	**69.89**	**57.39**
	+AdaptFormer (Chen et al., 2022a)	23.84	72.32	69.79	55.32
	+Affine-LN (Basu et al., 2024)	26.09	65.91	64.53	52.18
BB	Linear Probe	9.35	9.19	7.52	8.68
	Spatial Adapter	10.08	14.66	12.75	12.50

* Avg. of five: upper/lower lobe left, upper/lower lobe right, middle lobe.
† Avg. of five: myocardium, atrium/ventricle left, atrium/verticle right.
‡ Avg. of six: maximus left/right, medius left/right, minimus left/right.

Black-box methods are NOT competitive (Decoder Specialization)

Silva-Rodríguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MICCAI W-MedAGI'23 (Extension Under Review).

Zero-shot /Adaptation Oriented (3D Data)

FSEFT

Few-Shot Efficient Fine-Tuning

Transferability to novel tasks (new organs)

Setting	Method	Lung*	Heart†	Gluteus‡	Avg.
FULL	Fine-tuning (Tang et al., 2022)	19.59	53.14	55.37	42.70
	Fine-tuning (*Ours*)	**31.01**	60.79	65.35	52.38
PEFT	BitFit (Ben-Zaken et al., 2021)	14.79	48.90	39.43	34.28
	LoRA (Hu et al., 2022)	13.80	50.55	46.36	38.49
	AdaptFormer (Chen et al., 2022a)	18.82	53.35	48.61	40.26
	Affine-LN (Basu et al., 2024)	16.92	58.38	46.07	40.46
	Decoder fine-tuning	25.98	65.69	64.23	51.97
	+BitFit (Ben-Zaken et al., 2021)	26.17	65.78	64.34	52.10
	+LoRA (Hu et al., 2022)	26.16	**76.12**	**69.89**	**57.39**
	+AdaptFormer (Chen et al., 2022a)	23.84	72.32	69.79	55.32
	+Affine-LN (Basu et al., 2024)	26.09	65.91	64.53	52.18
BB	Linear Probe	9.35	9.19	7.52	8.68
	Spatial Adapter	10.08	14.66	12.75	12.50

* Avg. of five: upper/lower lobe left, upper/lower lobe right, middle lobe.
† Avg. of five: myocardium, atrium/ventricle left, atrium/verticle right.
‡ Avg. of six: maximus left/right, medius left/right, minimus left/right.

Additive PEFT outperform Selective methods

Silva-Rodríguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MICCAI W-MedAGI'23 (Extension Under Review).

Zero-shot /Adaptation Oriented (3D Data)

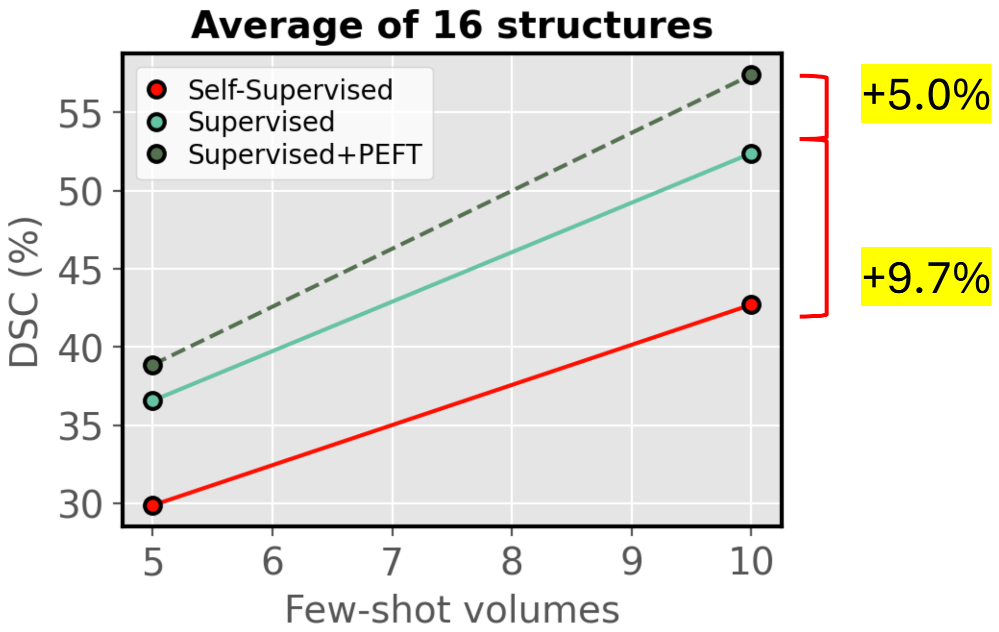
FSEFT

Few-Shot Efficient Fine-Tuning

Transferability to novel tasks (new organs)

Setting	Method	Lung*	Heart†	Gluteus‡	Avg.
FULL	Fine-tuning (Tang et al., 2022)	19.59	53.14	55.37	42.70
	Fine-tuning (*Ours*)	**31.01**	60.79	65.35	52.38
PEFT	BitFit (Ben-Zaken et al., 2021)	14.79	48.90	39.43	34.28
	LoRA (Hu et al., 2022)	13.80	50.55	46.36	38.49
	AdaptFormer (Chen et al., 2022a)	18.82	53.35	48.61	40.26
	Affine-LN (Basu et al., 2024)	16.92	58.38	46.07	40.46
	Decoder fine-tuning	25.98	65.69	64.23	51.97
	+BitFit (Ben-Zaken et al., 2021)	26.17	65.78	64.34	52.10
	+LoRA (Hu et al., 2022)	26.16	**76.12**	**69.89**	**57.39**
	+AdaptFormer (Chen et al., 2022a)	23.84	72.32	69.79	55.32
	+Affine-LN (Basu et al., 2024)	26.09	65.91	64.53	52.18
BB	Linear Probe	9.35	9.19	7.52	8.68
	Spatial Adapter	10.08	14.66	12.75	12.50

* Avg. of five: upper/lower lobe left, upper/lower lobe right, middle lobe.
† Avg. of five: myocardium, atrium/ventricle left, atrium/verticle right.
‡ Avg. of six: maximus left/right, medius left/right, minimus left/right.



Silva-Rodríguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MICCAI W-MedAGI'23 (Extension Under Review).

Zero-shot /Adaptation Oriented (3D Data)

FSEFT

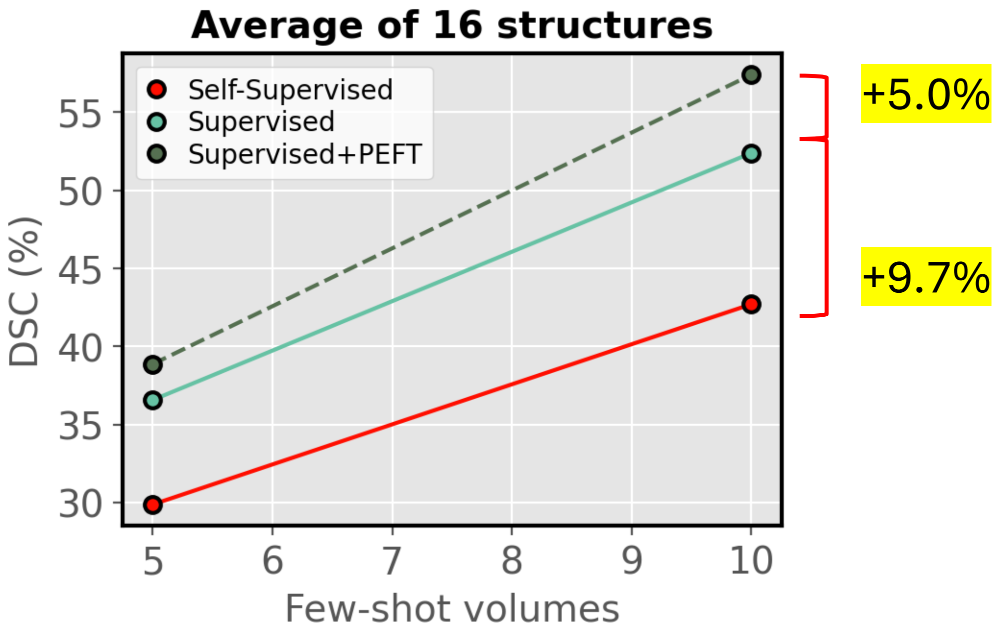
Few-Shot Efficient Fine-Tuning

Transferability to novel tasks (new organs)

*not for all structures

Setting	Method	Lung*	Heart†	Gluteus‡	Avg.
FULL	Fine-tuning (Tang et al., 2022)	19.59	53.14	55.37	42.70
	Fine-tuning (*Ours*)	**31.01**	60.79	65.35	52.38
PEFT	BitFit (Ben-Zaken et al., 2021)	14.79	48.90	39.43	34.28
	LoRA (Hu et al., 2022)	13.80	50.55	46.36	38.49
	AdaptFormer (Chen et al., 2022a)	18.82	53.35	48.61	40.26
	Affine-LN (Basu et al., 2024)	16.92	58.38	46.07	40.46
	Decoder fine-tuning	25.98	65.69	64.23	51.97
	+BitFit (Ben-Zaken et al., 2021)	26.17	65.78	64.34	52.10
	+LoRA (Hu et al., 2022)	26.16	**76.12**	**69.89**	**57.39**
	+AdaptFormer (Chen et al., 2022a)	23.84	72.32	69.79	55.32
	+Affine-LN (Basu et al., 2024)	26.09	65.91	64.53	52.18
BB	Linear Probe	9.35	9.19	7.52	8.68
	Spatial Adapter	10.08	14.66	12.75	12.50

* Avg. of five: upper/lower lobe left, upper/lower lobe right, middle lobe.
† Avg. of five: myocardium, atrium/ventricle left, atrium/verticle right.
‡ Avg. of six: maximus left/right, medius left/right, minimus left/right.



Silva-Rodríguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MICCAI W-MedAGI'23 (Extension Under Review).

Zero-shot /Adaptation Oriented (3D Data)

Challenges and future

Transferability between modalities, e.g. CT to MRI.

Model selection: we need to facilitate the adaptation/fine-tuning stage to practitioners.

How to know a priori if using black-box Adapters, or PEFT. Which PEFT method to use?

Improving PEFT for convolutional architectures.

Better benchmarks in generalist vs. specialized pre-training for 3D.

References

- Rebuffi et *al*. Learning Multiple Visual Domains with Residual Adapters. NeurIPS'17.
- Chen et *al*. Med3D: Transfer Learning for 3D Medical Image Analysis. ArXiv'19
- Cai et *al*. TinyTL: Reduce Memory, Not Parameters for Efficient On-Device Learning. NeurIPS'20.
- Frankle et *al*. Training BatchNorm and Only BatchNorm: On the Expressive Power of Random Features in CNNs. ICLR'21.
- Zhou et *al*. Model Genesis. MedIA'21.
- Ben-Zaken et *al*. BitFit: Simple Parameter-Efficient Fine-Tuning for Transformer-based Masked Language Models. ACL'22.
- Hu et *al*. LoRA: Low-Rank Adaptation of Large Language Models. ICLR'22.
- Chen et *al*. AdaptFormer Adapting Vision Transformers for Scalable Visual Recognition. NeurIPS'22.
- Tang et *al*. Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis. CVPR'22.
- Xie et *al*. UniMiSS: Universal Medical Self-Supervised Learning via Breaking Dimensionality Barrier. ECCV'22.
- Liu et *al*. CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection. ICCV'23.
- Ulrich et *al*. MultiTalent: A Multi-Dataset Approach to Medical Image Segmentation. MICCAI'23.
- Ye et *al*. UniSeg: A Prompt-driven Universal Segmentation Model as well as A Strong Representation Learner. MICCAI'23.
- Silva-Rodríguez et *al*. Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation. MICCAI W-MedAGI'23.
- Butoi et *al*. Universeg: Universal medical image segmentation. ICCV'23.
- Kirillov et *al*. Segment Anything. ICCV'23.
- Gao et *al*. Training Like a Medical Resident: Context-Prior Learning Toward Universal Medical Image Segmentation. CVPR'24.
- Li et *al*. How Well Do Supervised 3D Models Transfer to Medical Imaging Tasks?. ICLR'24.
- Liu et *al*. Universal and Extensible Language-Vision Models for Organ Segmentation and Tumor Detection from Abdominal CT. MedIA'24.
- Wang et *al*. SAM-Med3D: Towards General-Purpose Segmentation Models for Volumetric Medical Images. ArXiv'24.
- Gong et *al*. 3DSAM-adapter: Holistic Adaptation of SAM from 2D to 3D for Promptable Medical Image Segmentation. MedIA'24.
- Chen et *al*. MA-SAM: Modality-agnostic SAM Adaptation for 3D Medical Image Segmentation. MedIA'24.
- Ma et *al*. Segment Anything in Medical Images. Nat.Com.'24.
- Kulkarni et *al*. Anytime, Anywhere, Anyone: Investigating the Feasibility of SAM for Crowd-Sourcing Medical Image Annotations. MIDL'24.
- Huang et *al*. On The Challenges And Perspectives of Foundation Models For Medical Image Analysis. MedIA'24.
- Li et *al*. AdbomenAtlas: A Large Scale Detailed Annotated and Multi Center Dataset for Efficient Transfer Learning and Open Algorithmic Benchmarking. MedIA'24.
- Rakic et *al*. Tyche: Stochastic In-Context Learning for Medical Image Segmentation. CVPR'24.
- Basu et *al*. Strong Baselines for Parameter-Efficient Few-Shot Fine-Tuning-.. AAAI'24.
- Silva-Rodríguez et al. A Foundation Language-Image Model of the Retina: Encoding Expert Knowledge in Text Supervision. MedIA'24.
- Undandarao et al. No Zero-Shot without Exponential Data: Pretraining Concept frequency Determines Multimodal Model Performance. ICLRW-FM'24.