

Early Cancer Detection by Computed Tomography and Artificial Intelligence

Zongwei Zhou, PhD

Dept. Computer Science, Johns Hopkins University

zzhou82@jh.edu | (480)738-2575



JOHNS HOPKINS
UNIVERSITY

FELIX
Lustgarten
Foundation

2018



B. Vogelstein



A. Yuille



E. Fishman

FELIX-Civitas
Lustgarten &
McGovern
Foundation

2023



A. Yuille



C. Tomasetti

FELIX-Civitas
Lustgarten
Foundation &
NIH

2025



A. Yuille



Z. Zhou



K. Wang



Y. Yang



Milestone (2018 – 2023)

Detect early-stage cancer in the pancreas.

Milestone (2023 – 2025)

Detect cancer in the pancreas earlier than radiologists.

Milestone (2025 -)

Detect early-stage cancer in other organs.

Background

- Pancreatic cancer is extremely dangerous. Early detection is critical to enable effective treatment.
- Small tumors (diameter <2 cm) are very subtle and easy for skilled radiologists to overlook.
- Can AI help?
- (I) As a tool to assist radiologists?
- (II) As a technique to perform opportunistic screening of the general population?
To check the 80 millions (plus) CT scans taken yearly.

The Promise of AI

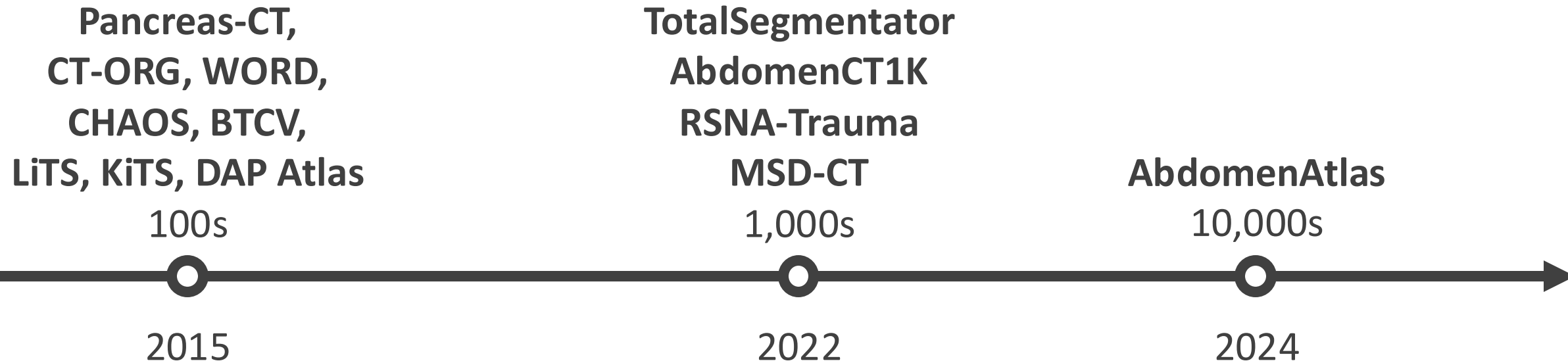
- AI has several advantages compared to humans.
- (I) AI can directly access the 3D voxels in CT scans while radiologists can only visualize 2D slices.
- (II) AI algorithms do not get tired or suffer from attentional effects.
- (III) AI algorithms can be trained on many more CT scans than a radiologist can see in a lifetime (est. 200,000 - 450,000 scans).

Challenges of Getting Big Data

- AI algorithms need data for training and testing. Assembling medical data is challenging.
- This requires accessing data from multiple institutions. This includes CT scans, radiology reports, and pathology reports.
- These scans must be cleaned, organized, and annotated.
- Annotation: Strong annotation requires voxel-wise annotation of scans. This is very time consuming (FELIX project at JHU took 25 person years).

Challenges of Getting Big Data

- Annotated data is becoming increasingly available. From 100 CT scans, to 1,000 scans, to 10,000 scans, and beyond. *Many are not for cancer.*
- Strong annotation is limited but weak annotation is practical.



New Paradigm for AI Testing and Training

- Testing: AI algorithms need to be tested on large and diverse datasets to ensure that they work in real world settings. Testing requires only *weak annotations*. *E.g., reports, bounding boxes, etc.*
- Training: AI algorithms require some strongly annotated data (time consuming to obtain), but this can be supplemented by synthetic data. Active learning – human-in-the-loop – can perform strong annotation very quickly.
- *Note: the original paradigm (as in FELIX) was to train and test AI on strongly annotated datasets. No longer necessary.*

Part I. Which AI Algorithms?

Part II. How to Annotate Data?

Part III. Can AI Find Early Cancer?

Part I. Which AI Algorithms?

- AI is an extremely dynamic research field. Novel AI algorithms are continually being created and improved.
- 99.99% of papers claim their AI is the best.
- But most of these algorithms do not perform well in open challenges.
- Very few are actually used in real-world clinical practice
- Design evaluation before design algorithms.

A Touchstone of Medical Segmentation

- We released a new standard for evaluating medical AI algorithms to promote fairness and reduce bias (Bassi et al., NeurIPS 2024).



Touchstone Benchmark: Are We on the Right Way for Evaluating AI Algorithms for Medical Segmentation?

Pedro R. A. S. Bassi^{1,2,3*} Wenxuan Li^{1*} Yucheng Tang⁴ Fabian Isensee^{5,6}
Zifu Wang⁷ Jieneng Chen¹ Yu-Cheng Chou¹ Saikat Roy^{5,8} Yannick Kirchhoff^{5,8,9}
Maximilian Rokuss^{5,8} Ziyang Huang¹⁰ Jin Ye¹¹ Junjun He¹¹ Tassilo Wald^{5,6}
Constantin Ulrich⁵ Michael Baumgartner^{5,6} Klaus H. Maier-Hein^{5,12} Paul Jaeger^{6,13}
Yiwen Ye¹⁴ Yutong Xie¹⁵ Jianpeng Zhang¹⁶ Ziyang Chen¹⁴ Yong Xia¹⁴
Zhaohu Xing¹⁷ Lei Zhu^{17, 18} Yousef Sadegheih¹⁹ Afshin Bozorgpour¹⁹
Pratibha Kumari¹⁹ Reza Azad²⁰ Dorit Merhof^{19,21} Pengcheng Shi²²
Ting Ma²² Yuxin Du²³ Fan Bai^{23,24} Tiejun Huang^{23,25} Bo Zhao^{10,23}
Haonan Wang¹⁸ Xiaomeng Li¹⁸ Hanxue Gu²⁶ Haoyu Dong²⁶
Jichen Yang²⁶ Maciej A. Mazurowski²⁶ Saumya Gupta²⁷ Linshan Wu¹⁸
Jiaxin Zhuang¹⁸ Hao Chen²⁸ Holger Roth⁴ Daguang Xu⁴
Matthew B. Blaschko⁷ Sergio Decherchi²⁹ Andrea Cavalli^{2,29,30}
Alan L. Yuille^{1†} Zongwei Zhou^{1†}

¹Department of Computer Science, Johns Hopkins University

²Department of Pharmacy and Biotechnology, University of Bologna

³Center for Biomolecular Nanotechnologies, Istituto Italiano di Tecnologia

⁴NVIDIA

⁵Division of Medical Image Computing, German Cancer Research Center (DKFZ)

⁶Helmholtz Imaging, German Cancer Research Center (DKFZ)

Full affiliations are given in Appendix F.

Code, Models & Data: <https://github.com/MrGiovanni/Touchstone>

A Touchstone of Medical Segmentation

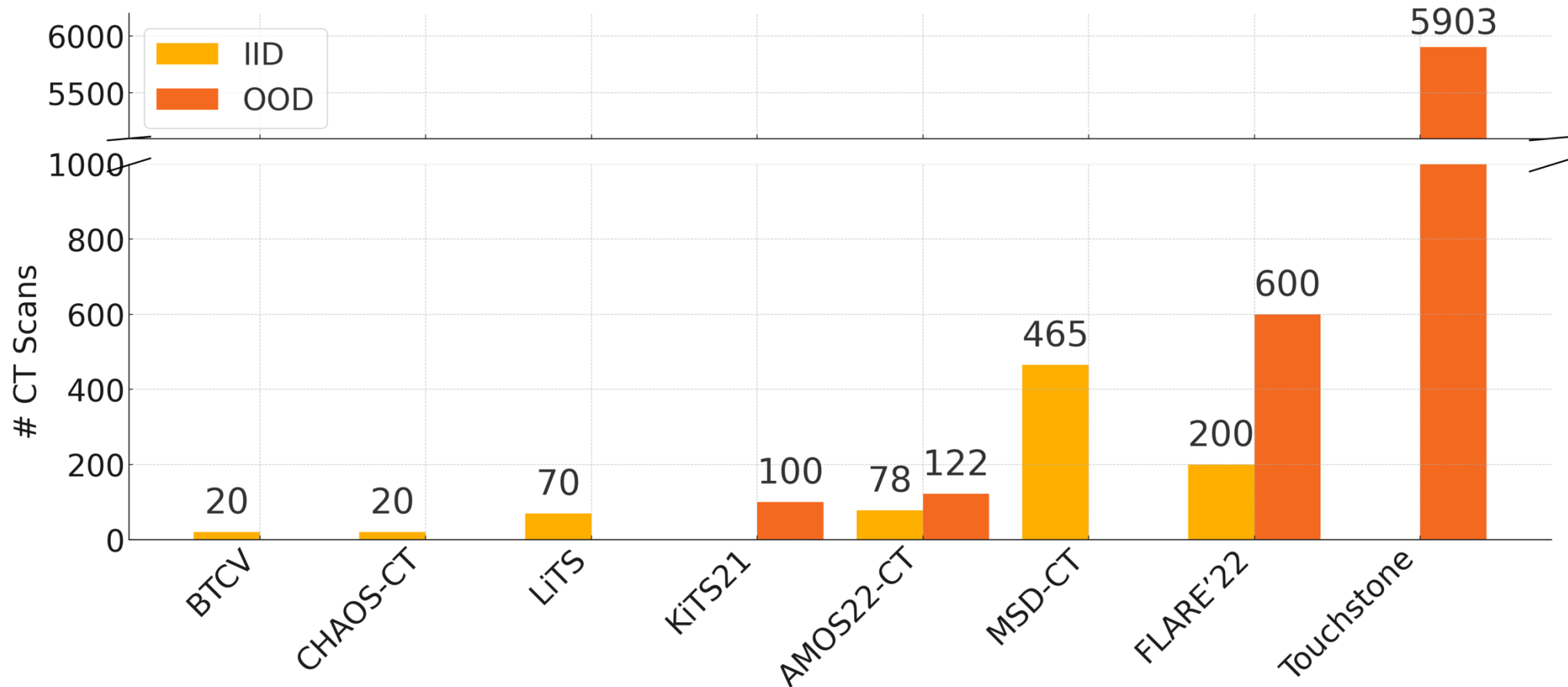
- (I) Inviting inventors to train their own algorithms.

Backbone	Lead Author	Institution	Publication	Backbone	Lead Author	Institution	Publication
U-Net	O. Ronneberger	Uni Freiburg	MICCAI	nnU-Net	Fabian Isensee	DKFZ	Nat. Methods
SegVol	Yuxin Du	SJTU	NeurIPS	MedFormer	Yunhe Gao	Rutgers	arXiv
CoTr	Yutong Xie	NPU	MICCAI	Swin UNETR	Ali Hatamizadeh	NVIDIA	MICCAIW
UniverSeg	Victor Ion Butoi	MIT	ICCV	UniSeg	Yiwen Ye	NPU	MICCAI
UNet++	Zongwei Zhou	ASU	TMI	MedNeXt	Saikat Roy	DKFZ	MICCAI
TransUNet	Jieneng Chen	JHU	ICMLW	MedSegDiff	Junde Wu	NUS	AAAI
Swin-Unet	Hu Cao	Huawei	ECCVW	3D UNeXt	Jeya Maria Jose	JHU	MICCAI
DiNTS	Yufan He	JHU	CVPR			

So far, **53 groups** have confirmed the contribution—we will invite more inventors of famous backbones for segmentation.

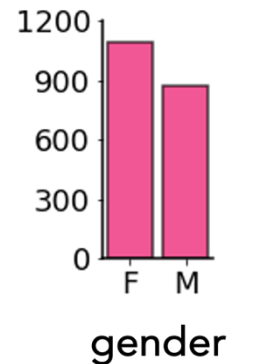
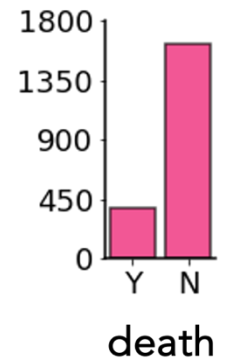
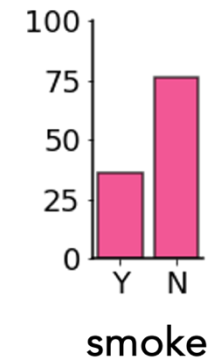
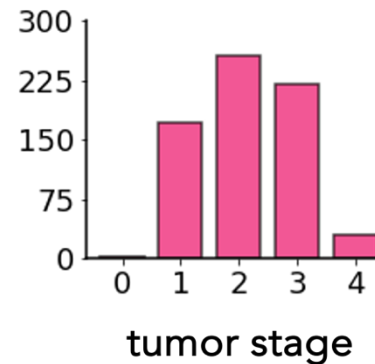
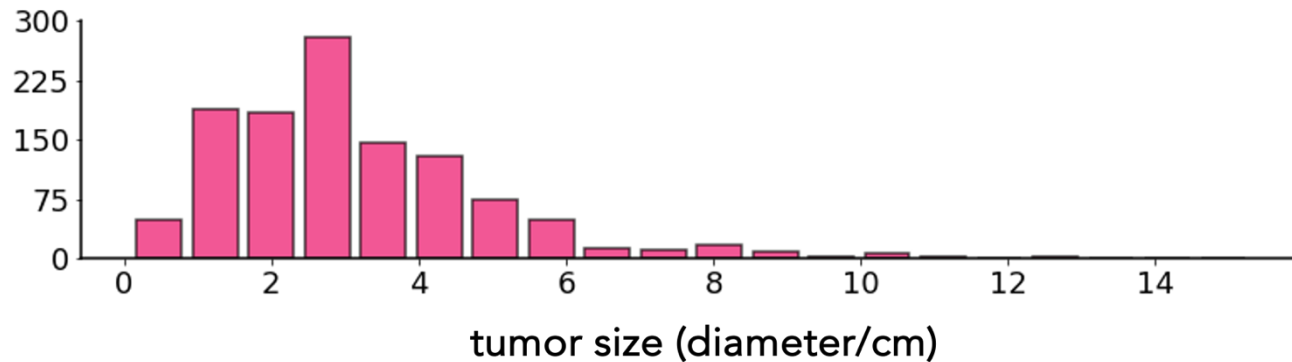
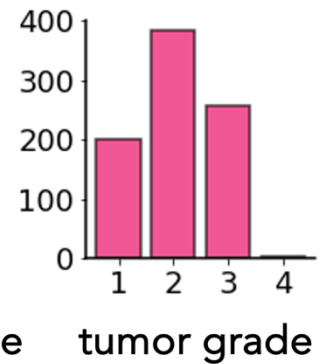
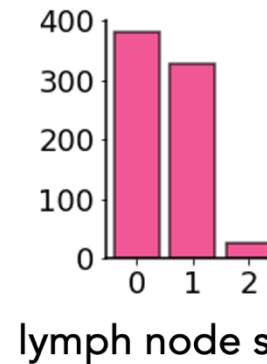
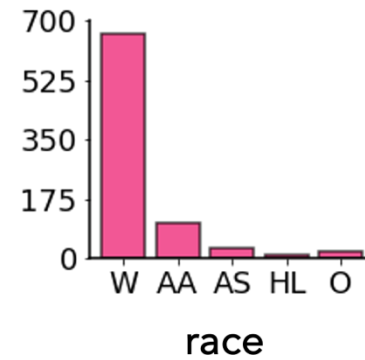
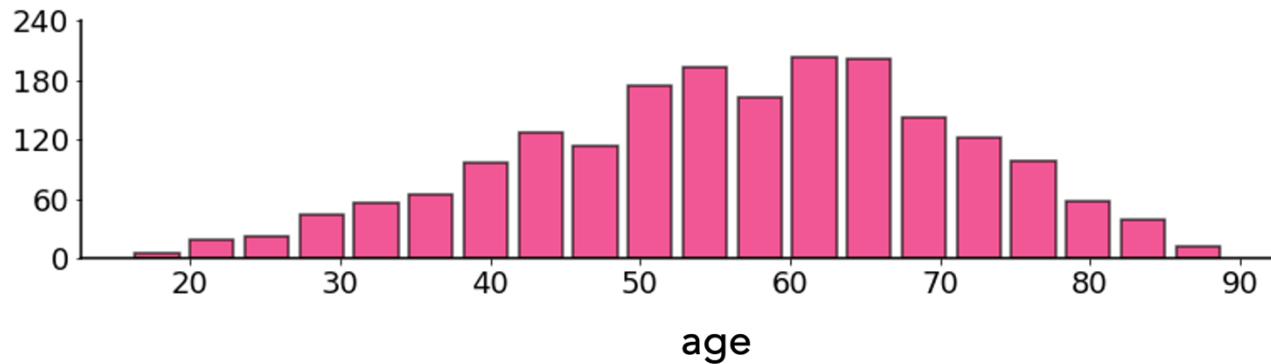
A Touchstone of Medical Segmentation















- (II) Third-party evaluating on a large ($n = 5,903$) dataset (i.e., FELIX).



A Touchstone of Medical Segmentation

- (III) Comprehensive metadata analysis



rank	model	average DSC	parameter	infer. speed	download
	MedNeXt	89.2	61.8M	★☆☆☆☆	
	MedFormer	89.0	38.5M	★★★★☆	
3	STU-Net-B	89.0	58.3M	★★☆☆☆	 checkpoint 
4	nnU-Net U-Net	88.9	102.0M	★★★★☆	 checkpoint 
5	nnU-Net ResEncL	88.8	102.0M	★★★★☆	 checkpoint 
6	UniSeg	88.8	31.0M	☆☆☆☆☆	
7	Diff-UNet	88.5	434.0M	★★★★☆	
8	LHU-Net	88.0	8.6M	★★★★★	 checkpoint 
9	NexToU	87.8	81.9M	★★★★☆	 checkpoint 
10	SegVol	87.1	181.0M	★★★★☆	 checkpoint 
11	U-Net & CLIP	87.1	19.1M	★★★★☆	

<https://github.com/MrGiovanni/Touchstone>

A Touchstone of Medical Segmentation

- We released a new standard for evaluating medical AI algorithms to promote fairness and reduce bias (Bassi et al., NeurIPS 2024).
- Large, out-of-distribution test set ($n = 5,903$).
- Large, multicenter training set ($n = 5,196$, from 76 centers).
- Inventor-involved training, third-party evaluation.
- Long-term investigation (Transformers, Mamba, newer architectures).



[GitHub.com/MrGiovanni/Touchstone](https://github.com/MrGiovanni/Touchstone)

A Touchstone of Medical Segmentation

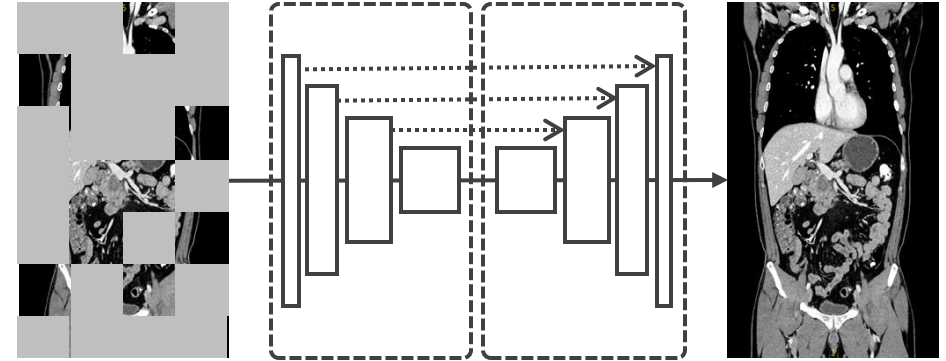
- We released a new standard for evaluating medical AI algorithms to promote fairness and reduce bias (Bassi et al., NeurIPS 2024).
- Large, out-of-distribution test set ($n = 5,903$).
- Large, multicenter training set ($n = 5,196$, from 76 centers).
- Inventor-involved training, third-party evaluation.
- Long-term investigation (Transformers, Mamba, newer architectures).
- **New** We have launched Touchstone 2.0 for benchmarking cancer tasks.



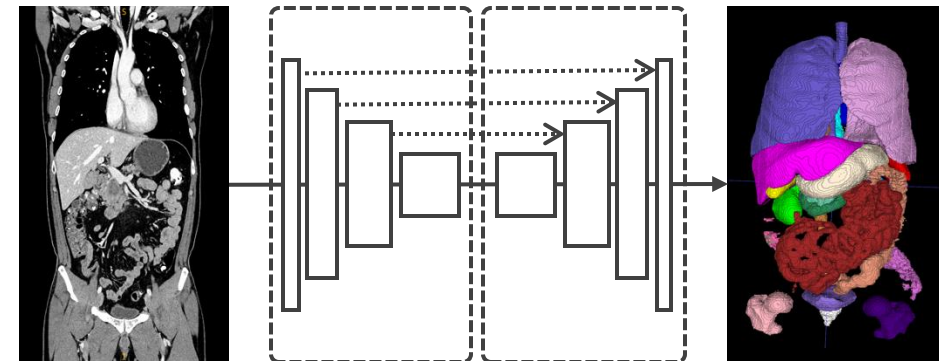
[GitHub.com/MrGiovanni/Touchstone](https://github.com/MrGiovanni/Touchstone)

A Touchstone of Foundation Models

- How well do supervised 3D models transfer to medical imaging tasks? (Li et al., ICLR 2024, Oral)
- The models are pre-trained on **9,000+** voxel-wise annotated 3D CT scans.
- The dataset & annotation used for training have been made public (Li et al., MEDIA 2024).

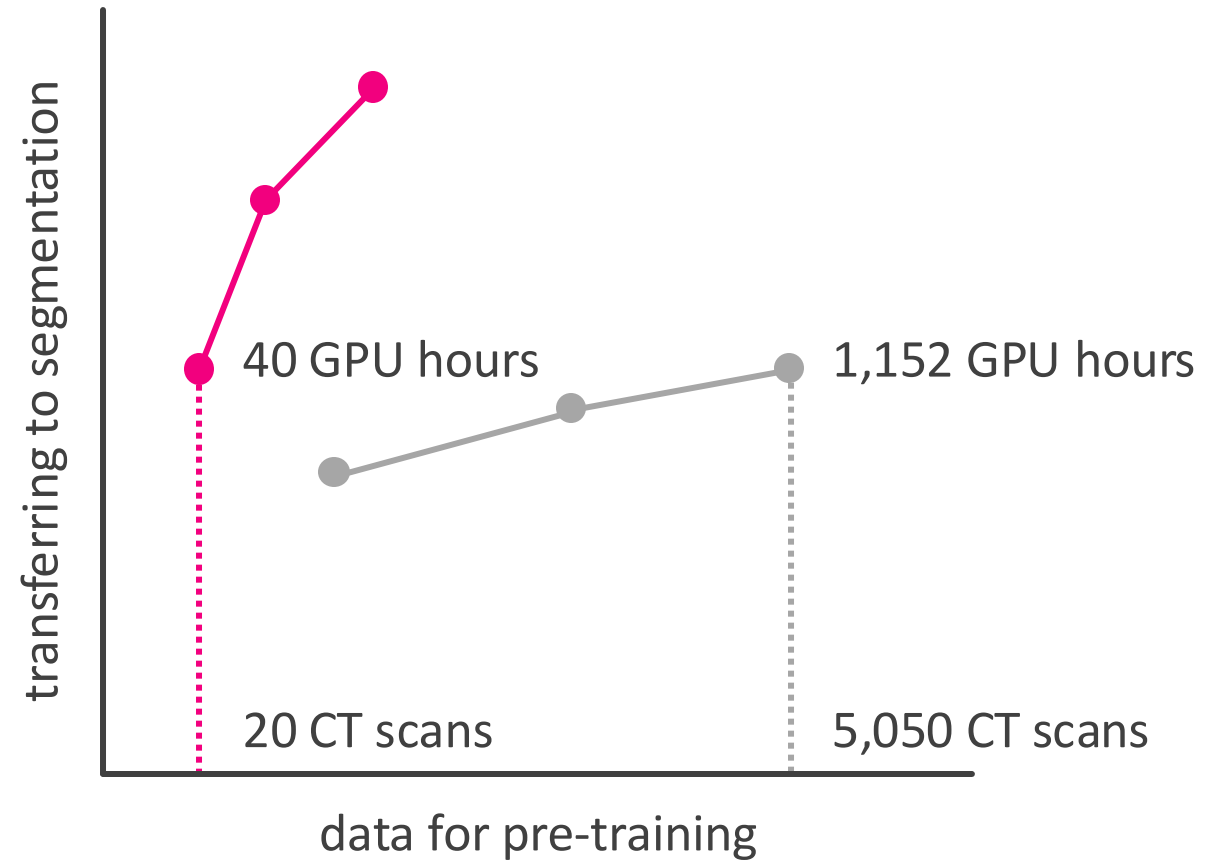


Self Supervision vs. (Full) Supervision



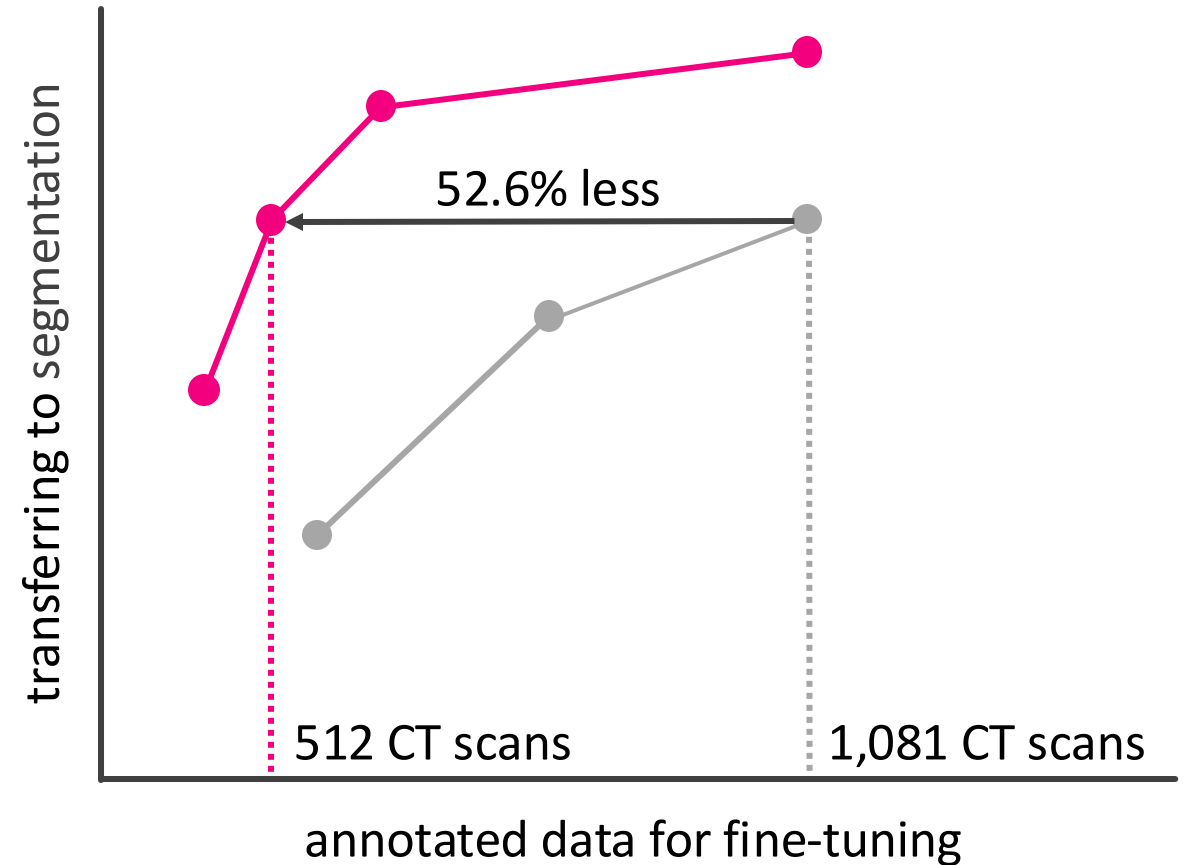
A Touchstone of Foundation Models

- (I) Supervised pre-training is more efficient in data and computation because of its clearly defined learning objective (e.g., segmentation).
- GPU hours reduced by **96.5%**.
- CT scans reduced by **99.6%**.



A Touchstone of Foundation Models

- (II) Supervised pre-training helps the model to learn image features that are relevant to downstream tasks (e.g., segmentation).
- Annotated data reduced by **52.6%**.



A Touchstone of Foundation Models

- How well do supervised 3D models transfer to medical imaging tasks? (Li et al., ICLR 2024, Oral)
- Supervised pre-training offers greater computation (**96.5%↑**), data (**99.6%↑**) and annotation (**52.6%↑**) efficiency.



[GitHub.com/MrGiovanni/SuPreM](https://github.com/MrGiovanni/SuPreM)

▼ Swin UNETR

name	params	pre-trained data	resources	download
Tang et al.	62.19M	5050 CT	Stars 1.1k	weights
Jose Valanaras et al.	62.19M	50000 CT/MRI	Stars 1.1k	weights
Universal Model	62.19M	2100 CT	Stars 639	weights
SuPreM	62.19M	2100 CT	ours 🌟	weights

▼ U-Net

name	params	pre-trained data	resources	download
Models Genesis	19.08M	623 CT	Stars 763	weights
UniMiSS	tiny	5022 CT&MRI	Stars 68	weights
	small	5022 CT&MRI		weights
Med3D	85.75M	1638 CT	Stars 2k	weights
DoDNet	17.29M	920 CT	Stars 184	weights
Universal Model	19.08M	2100 CT	Stars 639	weights
SuPreM	19.08M	2100 CT	ours 🌟	weights

Part I. Which AI Algorithms?

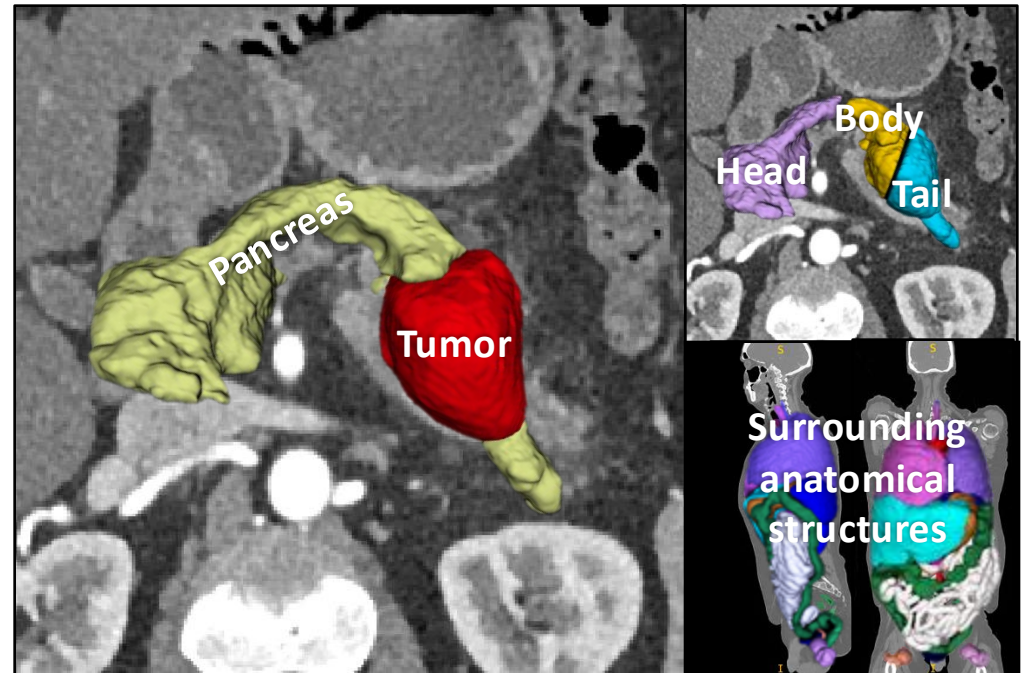
- AI is an extremely dynamic research field. Novel AI algorithms are continually being created and improved.
- We provide a testbed for rigorously benchmarking medical segmentation (Bassi et al., NeurIPS 2024), foundation models (Li et al., ICLR 2024, Oral), and visual question answering (Chen et al., Under Review).
- We developed effective/efficient algorithms based on the touchstone.
- (I) AI algorithms for detecting pancreatic tumors. **ePAI** (see later).
- (II) Foundational models for numerous downstream tasks. **SuPreM**

Part II. How to Annotate Data?

- Voxel-wise annotating tumors is very time consuming and requires experts (FELIX required 25 person years).
- Scaling this effort to all cancer types is not feasible.
- *What is the minimal amount of annotation needed to train an effective AI model for detecting a specific cancer type?*

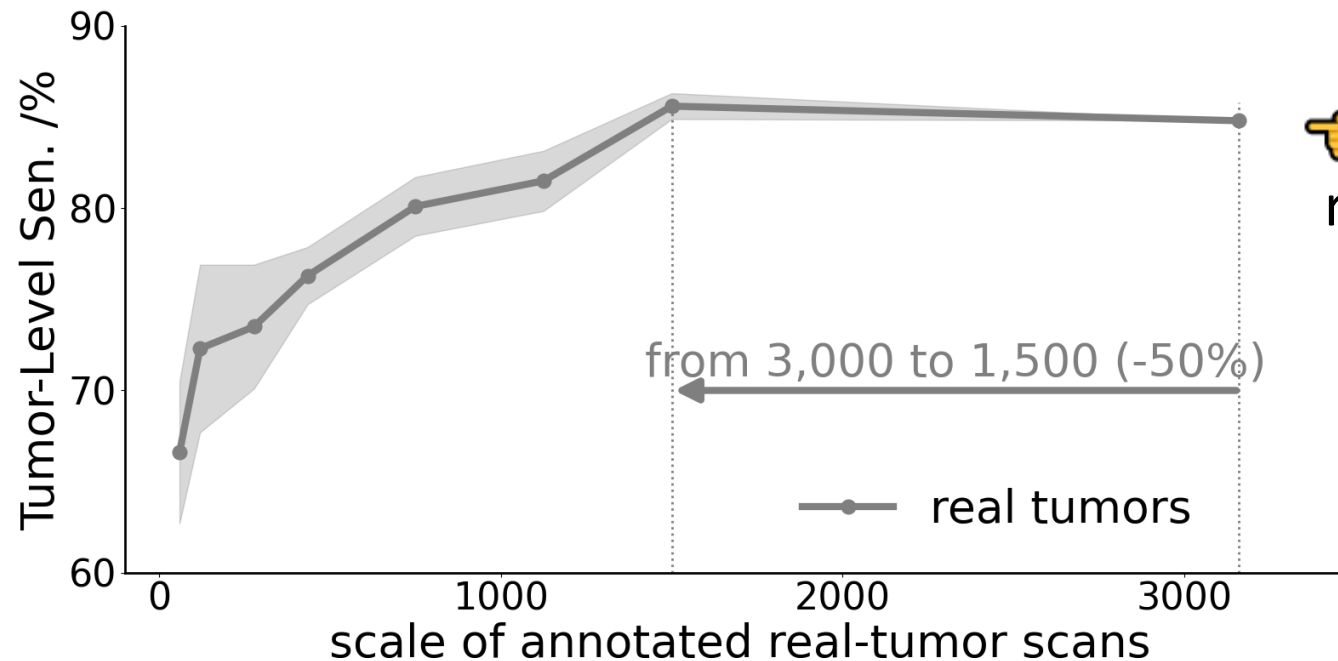
The FELIX Dataset (Private)

- >5,000 voxel-wise annotated CT scans, requiring **25** person years.
- It trains high-performance AI algorithms (Xia et al., medRxiv, 2022)
- Sensitivity = 97%, Specificity = 99%
- Comparable to radiologists.
- Generalizable to multiple centers.
- *But it is private!*
- *Are >5,000 necessary?*



Scaling Laws in Tumor Detection

- Detection performance reaches a plateau at $n = 1,500$ with more data.
- But annotating 1,500 CT scans remain challenging, can we do better?
- *Note: this is an internal validation!*

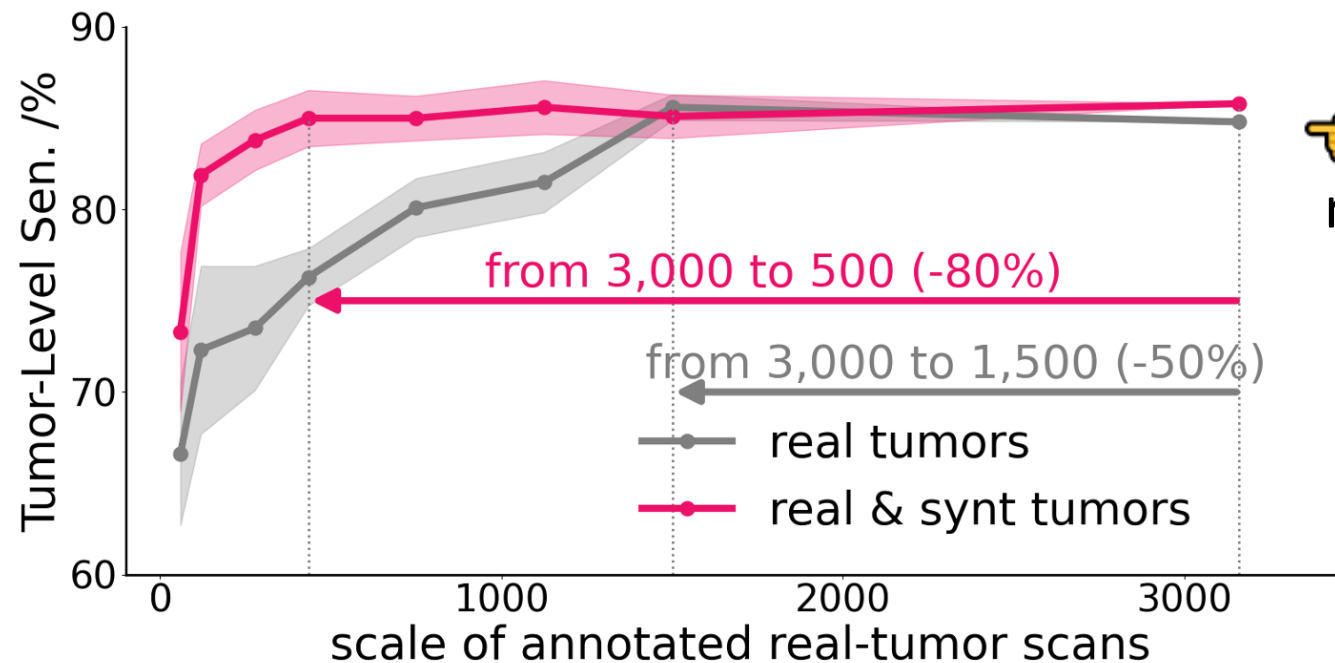


👉 comparable to radiologist-level performance



Scaling Laws in Tumor Detection

- Detection performance reaches a plateau at $n = 1,500$ with more data.
- Adding *Synthetic Data* achieves similar results using only $n = 500$ data.
- Synthetic data reduces the need for strongly annotated real data.

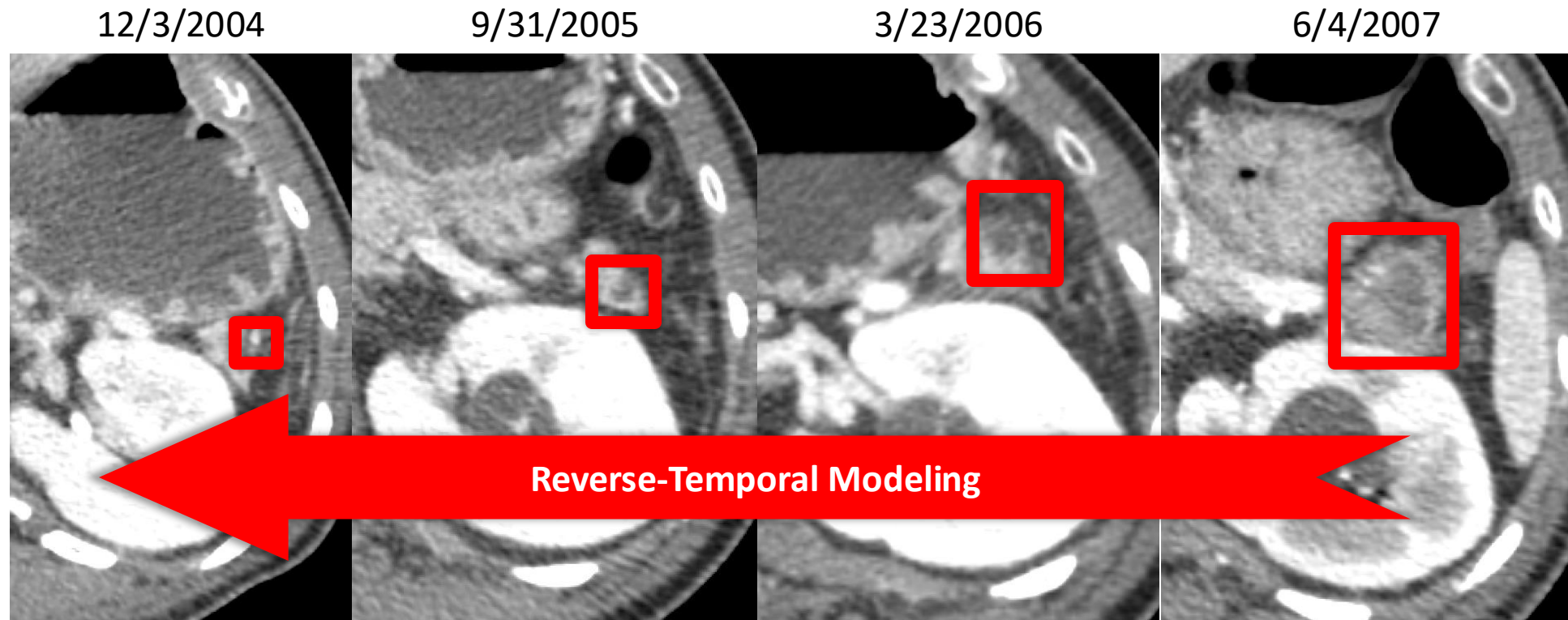


👉 comparable to radiologist-level performance



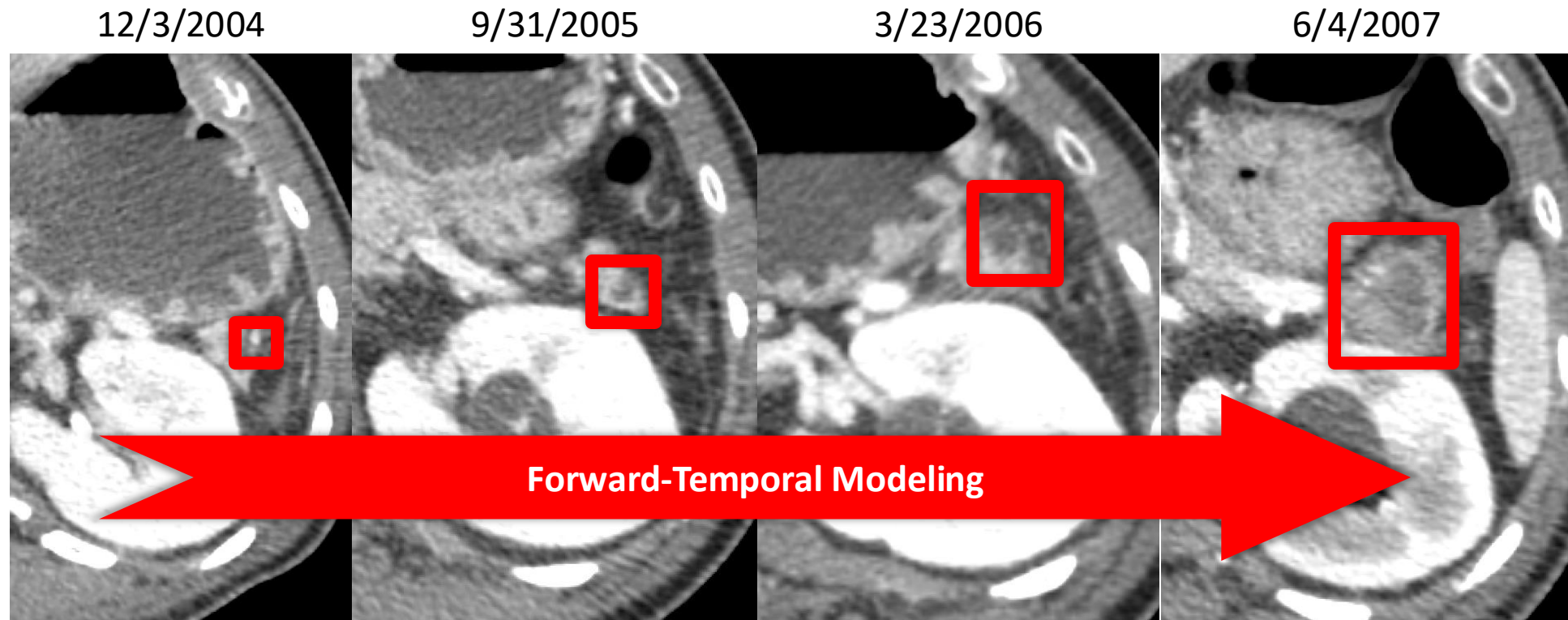
Synthetic Tumors as Time Machine

- Early-stage tumor scans are 10–20 times less common than late-stage scans in clinical datasets.



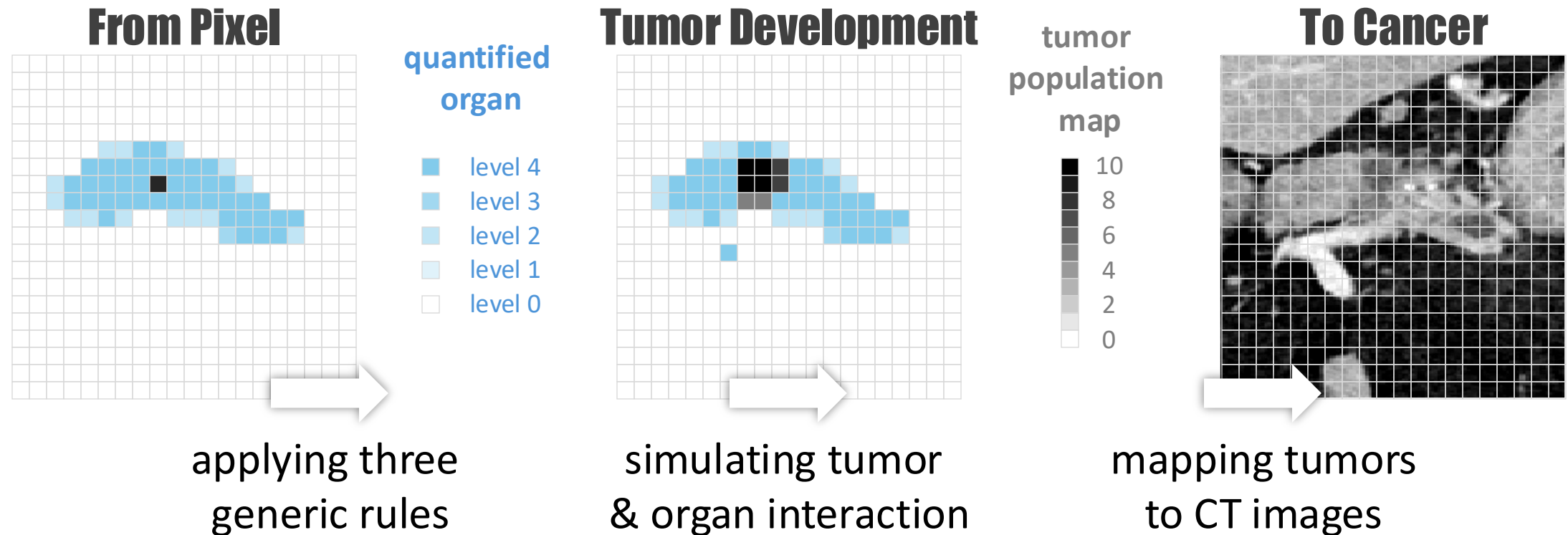
Synthetic Tumors as Time Machine

- Early-stage tumor scans are 10–20 times less common than late-stage scans in clinical datasets.



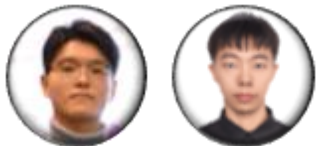
Synthetic Tumors as Time Machine

- We developed “game of life” to simulate tumor development ([Lai et al., MICCAI 2024](#)) and applied diffusion models to generate synthetic tumors ([Qi et al., CVPR 2024](#)).



Fake Data, Real Results

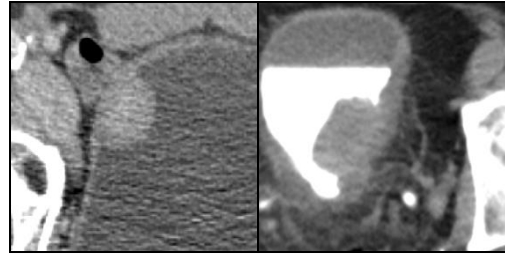
- A reverse-temporal generative model creates realistic early-stage CT scans from late-stage cases, helping AI learn small tumor patterns without needing many early-stage annotations.
- This improves detection sensitivity for small tumors (<2 cm) by **6%** compared to models trained only on real data.
- This cuts the need for annotations **from 1,500 to 500** without losing performance.
- *Note: 500 seems to be the minimal requirement for each type of cancer!*



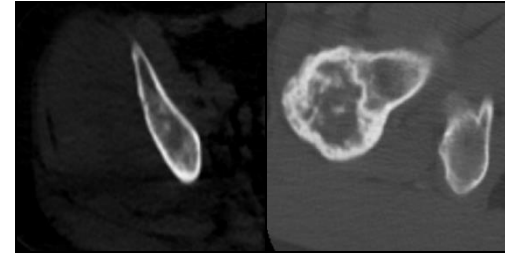
Synthetic Tumors in Different Organs



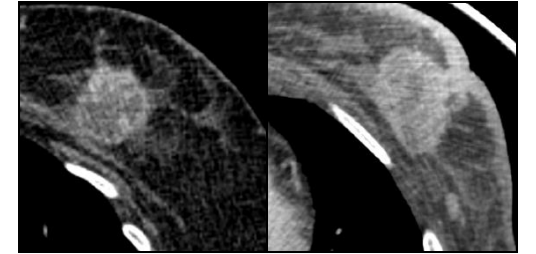
(a) real or fake test



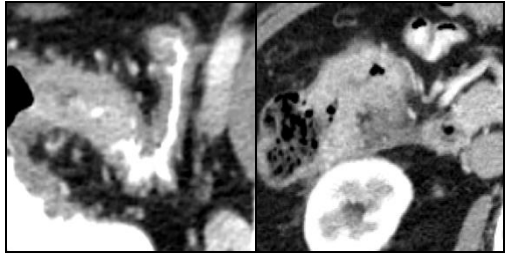
(b) bladder tumor



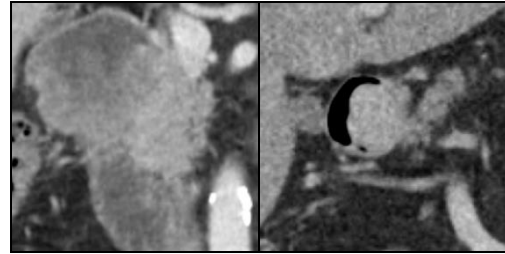
(c) bone tumor



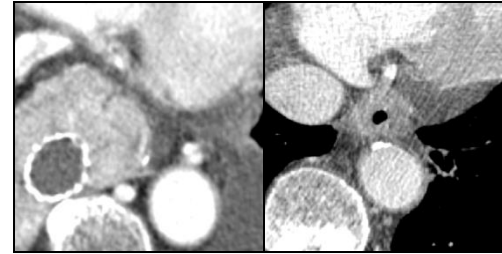
(d) breast tumor



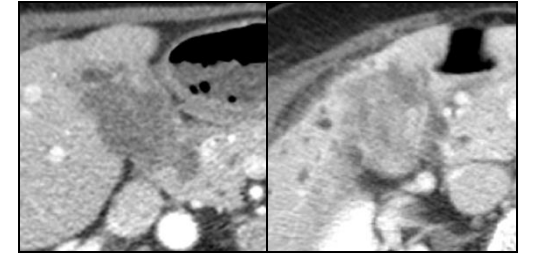
(e) colon tumor



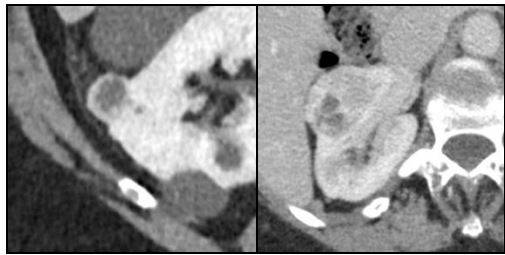
(f) duodenum tumor



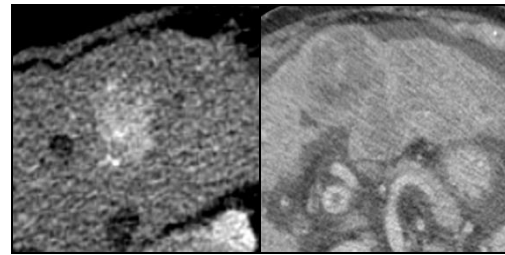
(g) esophagus tumor



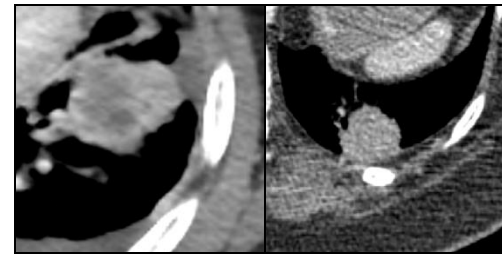
(h) gallbladder tumor



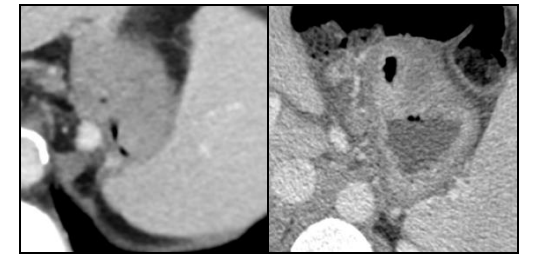
(i) kidney tumor



(j) liver tumor



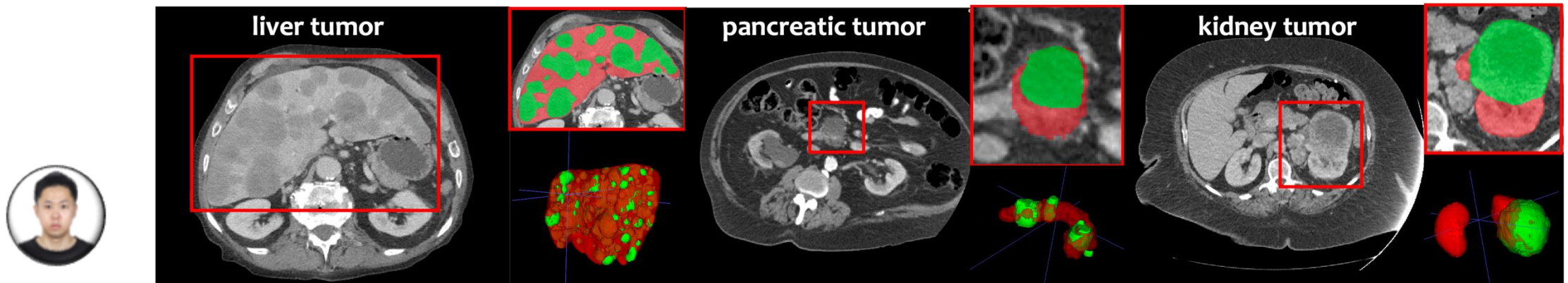
(k) lung tumor



(l) stomach tumor

Synthetic Tumors for Efficient Annotation

- Training on many synthetic tumors makes the AI highly sensitive, offering a strong starting point for radiologist review and edit at least **5× faster** (Zhou et al., ISBI 2024). *Active learning, human-in-the-loop.*
- (I) Editing an AI-generated tumor takes **~1 minute**.
- (II) Removing a false positive takes **<5 seconds**.
- In contrast, manual annotation from scratch takes **4–5 minutes**.



AbdomenAtlas 2.0 (Public)

- A large, multi-center (*real*) tumor dataset with per-voxel annotations.
- 9,262 CT scans + 8,562 tumor masks.
- Get early access 📌 <https://www.zongweiz.com/dataset>
- Much larger than existing public datasets.
- (I) **46x** LiTS (201 liver scans).
- (II) **22x** MSD-Pancreas (420 pancreatic scans).
- (III) **15x** KiTS (599 kidney scans).



A team of 23 board-certified radiologists



Part II. How to Annotate Data?

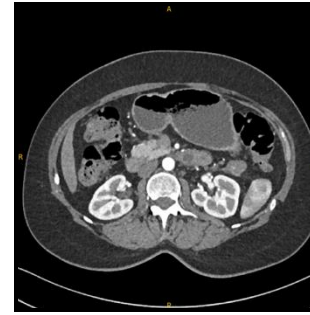
- Voxel-wise annotating tumors is very time consuming and requires experts (FELIX required 25 person years).
- Exploiting generative models to create realistic synthetic tumors. Radiologists check the realism of these synthetic tumors and provide feedback.
- Synthetic tumors supplement the strong per-voxel annotations in dataset like FELIX to provide additional training data.
- Synthetic tumors can be created at very small scale. This helps because current datasets are short of very small tumors (<2 cm, <1 cm) because they are so hard to detect.
- *Note: synthetic tumors are used for training AI only – not for testing.*

Part III. Can AI Find Early Cancer?

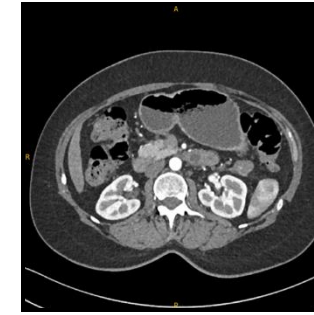
- We test the AI on small tumors – defined as <2 cm in diameter.
- The results are reported on JHU and on data from other institutions with weak annotation.
- Institutions and consortiums: UW (n = 2,828), UCSF (n = 1,176), CoH/TGen (n = 521), Heidelberg (n = 97).
- *Note: this project is ongoing with more data from new institutions and consortiums.*

ePAI

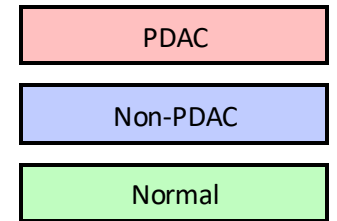
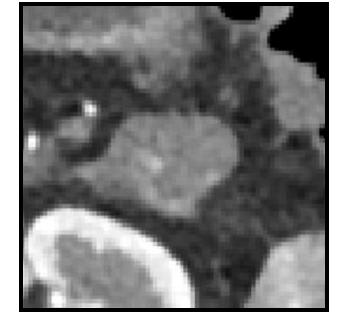
- ePAI: early Pancreatic cancer detection with Artificial Intelligence
- Its three stages exploited
- Part I. the best AI algorithm, & foundation models
- Part II. AbdomenAtlas, FELIX, & synthetic data



Stage 1
segmentation



Stage 2
localization



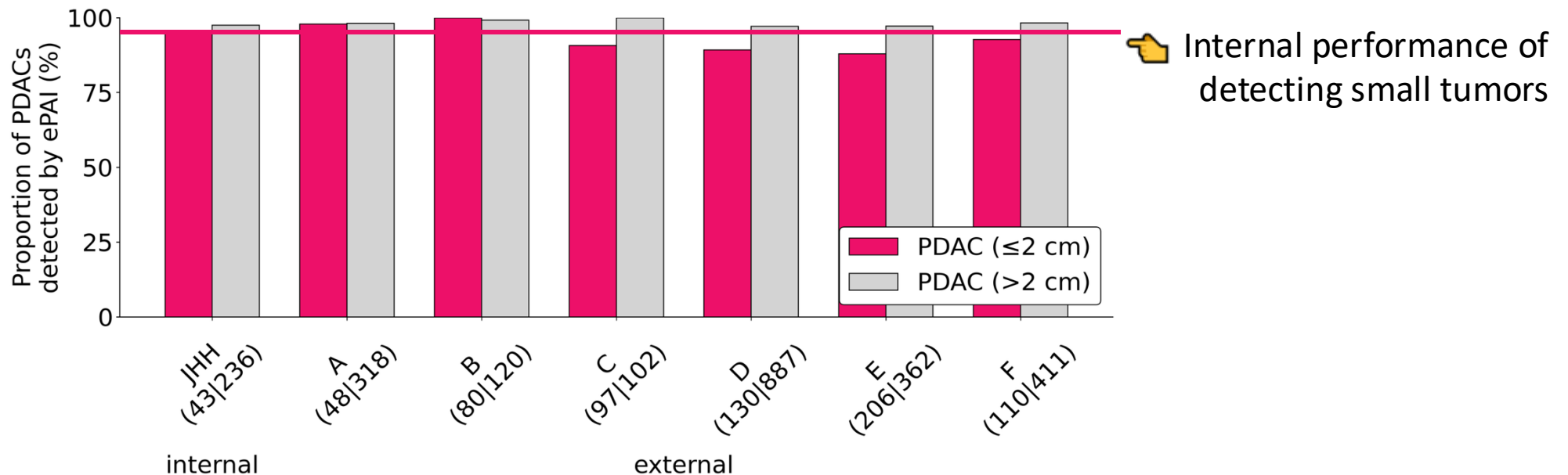
Stage 3
classification

Current Results

- Validation ePAI on early-tumor detection performance in:
- (I) Small tumors ($< 2\text{cm}$).
- (II) Prediagnostic scans.
- (III) Comparing with radiologists.

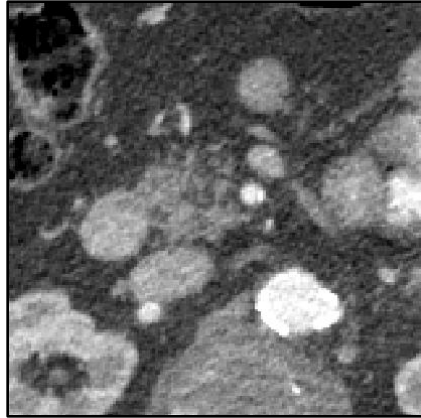
Current Results

- (I) Small tumors (< 2cm).
- Performance is particularly strong for small PDACs – sensitivity for (<2cm) tumors was 95% on FELIX, with a specificity of 98%.
- Performance slightly degrades for some out-of-domain testing.



Current Results

- (II) Prediagnostic scans.
- Definition



**Prediagnostic CT
scans**

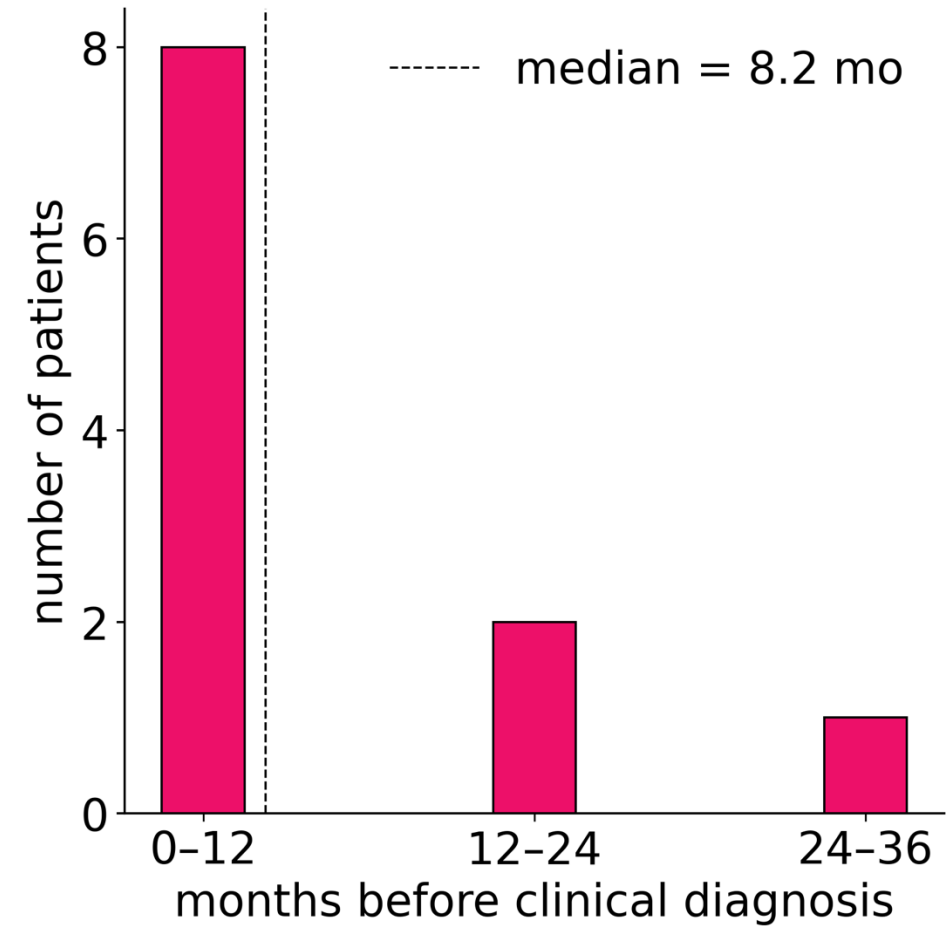


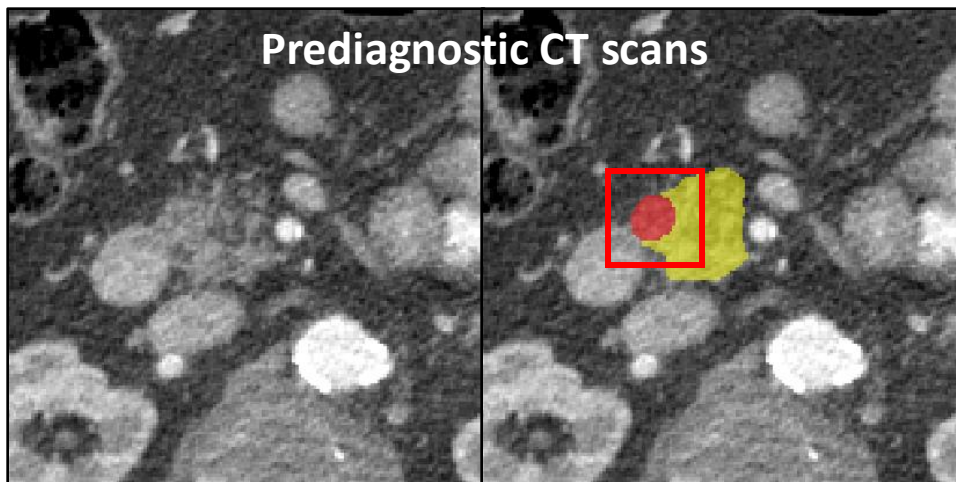
**Diagnostic CT
scans**



Current Results

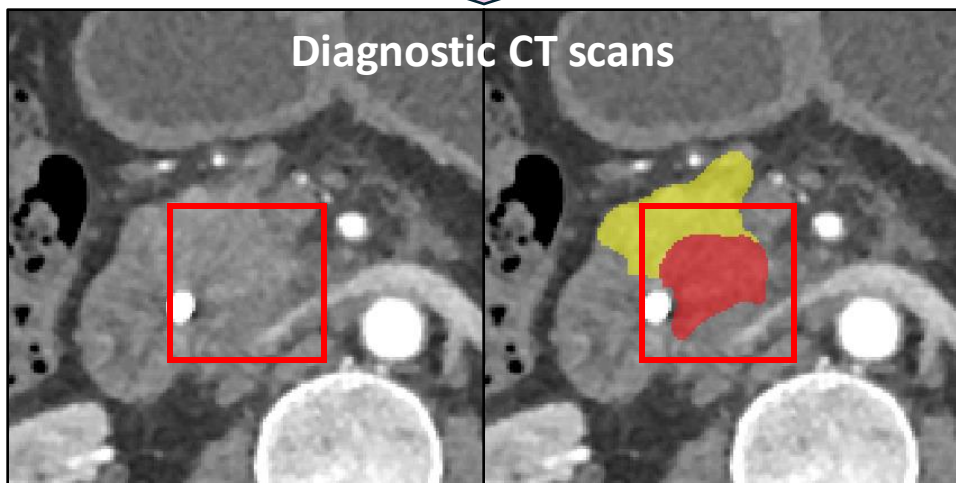
- (II) Prediagnostic scans.
- ePAI successfully detected PDAC in **36 of 58** patients (sensitivity = **62%**) that had been overlooked by radiologists, with a median lead time of **244** days before clinical diagnosis.





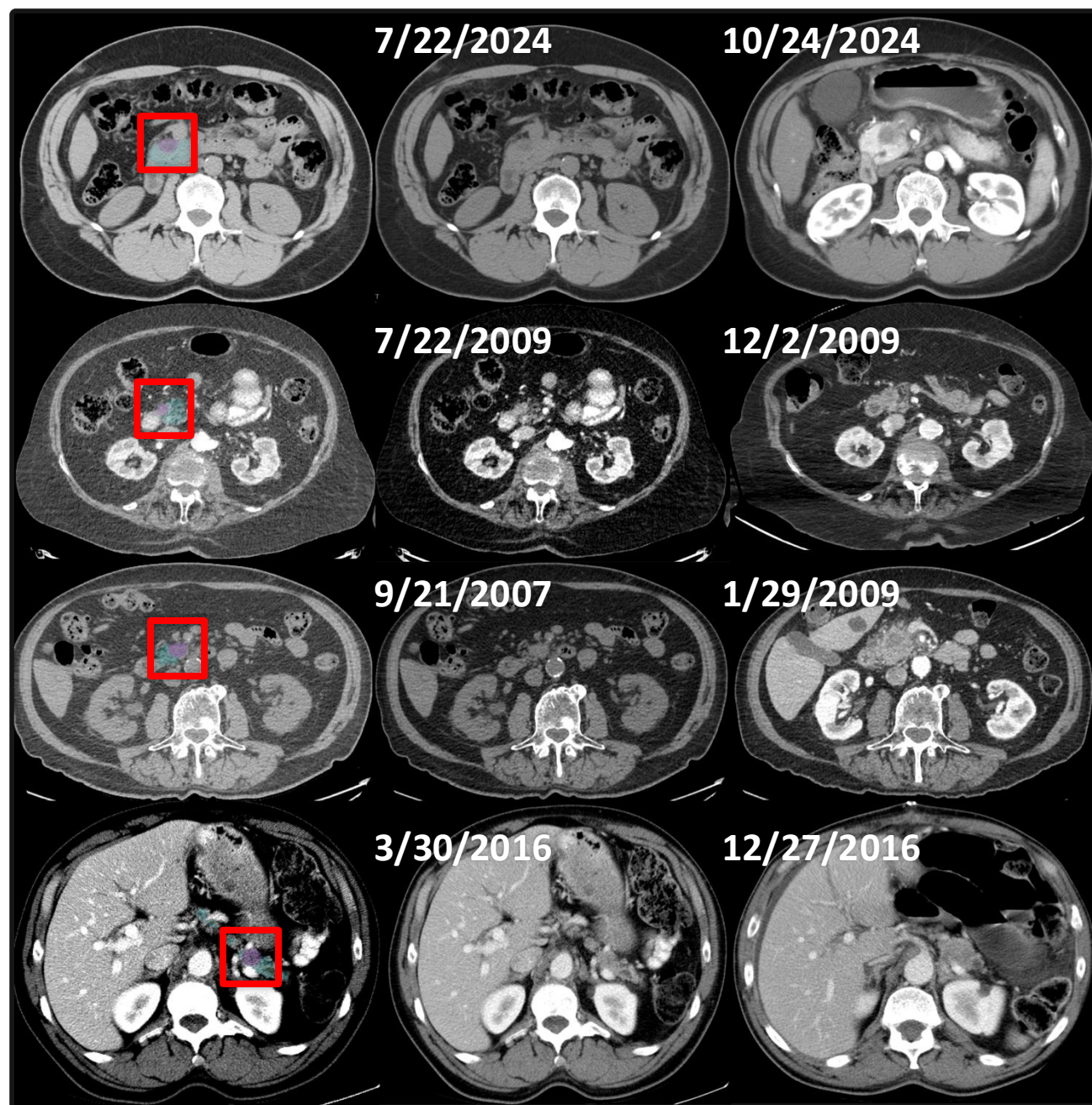
Radiologists

ePAI



Radiologists

ePAI



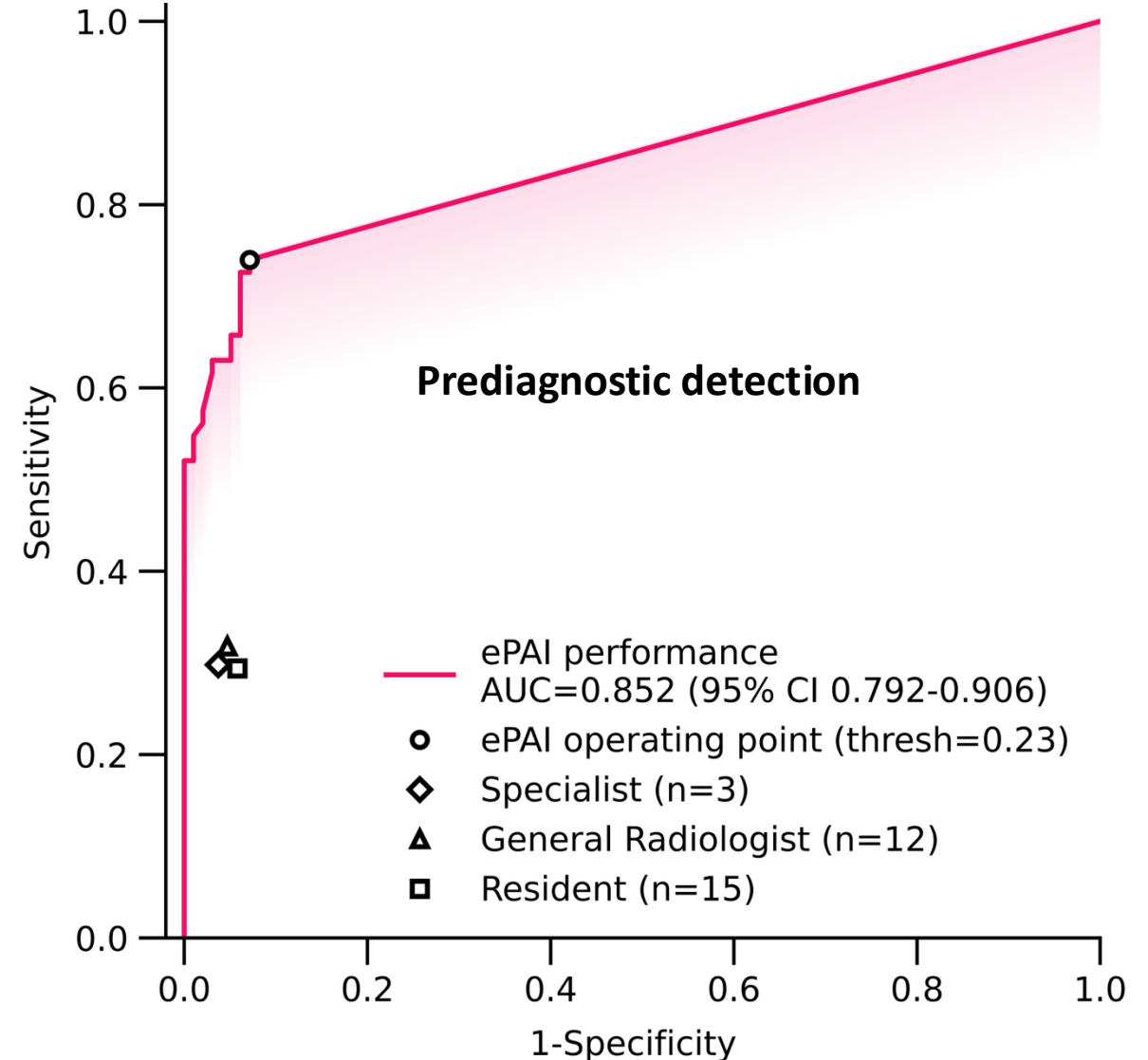
ePAI

Prediagnostic

Diagnostic

Current Results

- (III) Comparing with radiologists.
- Few studies report *radiologist performance* on these tasks. Existing studies show low sensitivity (30–40%) for detecting small tumors (<2 cm).
- In our ongoing study, a team of **30** radiologists achieved **34%** sensitivity and **94%** specificity.
- ePAI obtained **2x** sensitivity than radiologists with similar specificity.



Part III. Can AI Find Early Cancer?

- It is now being practical to create datasets which are sufficiently large for training and testing AI algorithms for early cancer detection.
- It is also possible to use AI algorithms to provide strong annotations for these datasets and to supplement them with synthetic data.
- Performance of the AI, particularly for detecting small ($< 2\text{cm}$) tumors is high, works on data from a variety of institutions/consortiums, and performs well compared with radiologists.
- *Is this enough for opportunistic perspective? At present, too many false positives considering the low disease prevalence (0.005 - 0.01%).*



Key Takeaways (Technical)

- Testing: AI algorithms need to be tested on large and diverse datasets to ensure that they work in real world settings. Testing requires only *weak annotations*. *E.g., reports, bounding boxes, etc.*
- Training: AI algorithms require some strongly annotated data (time consuming to obtain), but this can be supplemented by synthetic data. Active learning – human-in-the-loop – can perform strong annotation very quickly.
- *Note: the same approach is being applied to cancer in other organs.*
- Cancer Grand Challenges: AI-human collaborations in cancer (\$25M).

Reference

- Bassi, Pedro RAS, Wenxuan Li, ..., Alan Yuille, and Zongwei Zhou. "Touchstone benchmark: Are we on the right way for evaluating ai algorithms for medical segmentation?" *NeurIPS*, 2024.
- Chen, Qi, ..., Alan Yuille, and Zongwei Zhou. "Towards generalizable tumor synthesis." *CVPR*, 2024.
- Lai, Yuxiang, ..., Alan Yuille, and Zongwei Zhou. "From pixel to cancer: Cellular automata in computed tomography." *MICCAI*, 2024.
- Li, Wenxuan, Alan Yuille, and Zongwei Zhou. "How well do supervised 3d models transfer to medical imaging tasks?" *ICLR*, 2025 (oral).
- Li, Wenxuan, ..., Alan Yuille, and Zongwei Zhou. "Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking." *MEDIA*, 2024.
- Xia, Yingda, Qihang Yu, ..., Zongwei Zhou, ..., Alan Yuille. "The felix project: Deep networks to detect pancreatic neoplasms." *medRxiv*, 2022.
- Zhou, Xinze, ..., Alan Yuille, and Zongwei Zhou. "Efficient Human-in-the-Loop Pancreatic Tumor Annotation via Large-Scale Pre-Trained Model with Adaptive Post-Processing." *ISBI*, 2025.