Medical Image Analysis (2024)

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

AbdomenAtlas: A Large-Scale, Detailed-Annotated, & Multi-Center Dataset for Efficient Transfer Learning and Open Algorithmic Benchmarking

Wenxuan Li[a], Chongyu Qu[a], Xiaoxi Chen[b], Pedro R. A. S. Bassi[a,c,d], Yijia Shi[e], Yuxiang Lai[a,f], Qian Yu[g], Huimin Xue[h], Yixiong Chen[a], Xiaorui Lin[i], Yutong Tang[i], Yining Cao[i], Haoqi Han[i], Zheyuan Zhang[j], Jiawei Liu[j], Tiezheng Zhang[a], Yujiu Ma[k], Jincheng Wang[l], Guang Zhang[m,n,o], Alan Yuille[a], Zongwei Zhou[a,*]

[a] *Department of Computer Science, Johns Hopkins University*
[b] *Department of Bioengineering, University of Illinois Urbana-Champaign*
[c] *Alma Mater Studiorum - University of Bologna*
[d] *Center for Biomolecular Nanotechnologies, Istituto Italiano di Tecnologia*
[e] *LKS Faculty of Medicine, The University of Hong Kong*
[f] *Department of Computer Science, Southeast University*
[g] *Department of Radiology, Southeast University Zhongda Hospital*
[h] *Department of Medical Oncology, The First Hospital of China Medical University*
[i] *The Second Clinical College, China Medical University*
[j] *Department of Mechanical Engineering and the Laboratory of Computational Sensing and Robotics, Johns Hopkins University*
[k] *Center of Reproductive Medicine, Department of Obstetrics and Gynecology, Shengjing Hospital of China Medical University*
[l] *Radiology Department, the First Affiliated Hospital, School of Medicine, Zhejiang University*
[m] *Department of Health Management, The First Affiliated Hospital of Shandong First Medical University & Shandong Provincial Qianfoshan Hospital*
[n] *Shandong Engineering Research Center of Health Management*
[o] *Shandong Institute of Health Management*

ARTICLE INFO

Article history:
Received 1 May 2013
Received in final form 10 May 2013
Accepted 13 May 2013
Available online 15 May 2013

Communicated by S. Sarkar

2000 MSC: 41A05, 41A10, 65D05, 65D17

Keywords: Annotation, Dataset, Transfer Learning, Benchmark

ABSTRACT

We introduce the largest abdominal CT dataset (termed AbdomenAtlas) of 20,460 three-dimensional CT volumes sourced from 112 hospitals across diverse populations, geographies, and facilities. AbdomenAtlas provides 673K high-quality masks of anatomical structures in the abdominal region annotated by a team of 10 radiologists with the help of AI algorithms. We start by having expert radiologists manually annotate 22 anatomical structures in 5,246 CT volumes. Following this, a semi-automatic annotation procedure is performed for the remaining CT volumes, where radiologists revise the annotations predicted by AI, and in turn, AI improves its predictions by learning from revised annotations. Such a large-scale, detailed-annotated, and multi-center dataset is needed for two reasons. Firstly, AbdomenAtlas provides important resources for AI development at scale, branded as *large pre-trained models*, which can alleviate the annotation workload of expert radiologists to transfer to broader clinical applications. Secondly, AbdomenAtlas establishes a large-scale benchmark for evaluating AI algorithms—the more data we use to test the algorithms, the better we can guarantee reliable performance in complex clinical scenarios. An ISBI & MICCAI challenge named *BodyMaps: Towards 3D Atlas of Human Body* was launched using a subset of our AbdomenAtlas, aiming to stimulate AI innovation and to benchmark segmentation accuracy, inference efficiency, and domain generalizability. We hope our AbdomenAtlas can set the stage for larger-scale clinical trials and offer exceptional opportunities to practitioners in the medical imaging community. Codes, models, and datasets are available at https://www.zongweiz.com/dataset

1. Introduction

Large pre-trained models have revolutionized natural language processing (NLP) with examples like GPTs (Brown et al., 2020) and LLaMA (Touvron et al., 2023). However, the road map to achieve such transformative models remains unfolding in computer vision (CV) despite fervent explorations being undertaken. The current strategies in CV are diverse: using pixels only (e.g., LVM (Bai et al., 2023)), combining pixels with texts (e.g., LLaVA (Liu et al., 2023a)), or incorporating detailed human annotations (e.g., SAM (Kirillov et al., 2023)). While these strategies have shown promise, they have yet to match the success level of language models that can be widely applicable across target tasks. This variety reflects the inherent complexity and varied requirements in processing image data compared with text data (Zhang and Metaxas, 2023).

A consensus is that *large pre-trained models must be trained on massive, diverse datasets* (Moor et al., 2023; Blankemeier et al., 2024). The road we must take is to prepare massive, diverse datasets—and it would be even better if they were annotated. For language models, very large and diverse datasets are fairly easy to obtain (e.g., 250B web pages in the Common Crawl repository[1]). For vision models, we are still very far from having a data source of comparable size and diversity (e.g., 5.85B images in the LAION-5B dataset (Schuhmann et al., 2022)). In particular, medical vision, bearing some resemblance to computer vision, is relatively new in this exploration and pretty much a vacuum in the search for sizable datasets (Chen et al., 2022), especially for the most dominant 3D medical images (Blankemeier et al., 2024). Moreover, a unique issue of medical images extends to variations in data collections, imaging protocols, and patient demographics (McKinney et al., 2020; Singh et al., 2022), which is often overlooked in most existing datasets, as summarized in Table 1 and §2.1. This raises a pressing concern about the generalizability of the pre-trained models.

This paper does not intend to discuss how to create GPT-like vision models in medical imaging but endeavors to provide the required data and annotations that could catalyze such discussions. We have collected and annotated **20,460** CT volumes, totaling **673K** high-quality masks of anatomical structures in the abdominal region. These CT volumes are taken from **112** hospitals in **19** countries, making this effort unprecedented in scale. It is, by far, the most extensive annotated medical dataset for AI benchmark and promises to be a valuable asset for the development of large pre-trained models in the medical domain. We name this dataset AbdomenAtlas. A large dataset from diverse centers is needed for two main reasons: (I) The performance of AI algorithms is known to improve when they are trained on more data; the more data we use to test the algorithms, the better we can guarantee good performance under real-world conditions (e.g., clinical settings). (II) It is critically important to train and test AI algorithms on data from different centers because AI researchers have found that algorithms trained on data from one center may fail to generalize to data from other centers (as exemplified in DeGrave et al. (2021) and Geirhos et al. (2020)).

In the remainder of this paper, we begin with a review of the preexisting medical datasets that are publicly available and highlight the unique properties of our AbdomenAtlas in §2. We then describe in depth the construction of AbdomenAtlas in §3, elaborating time-consuming manual annotation for 5,246 CT volumes and efficient semi-automatic annotation procedure for the remaining 15,214 CT volumes. Following this, two practical applications of our AbdomenAtlas are presented. Firstly, §4 introduces a suite of large pre-trained models (SuPreM) enabling efficient transfer learning across numerous downstream tasks, with a special analysis on transfer learning efficiency and ability. Secondly, §5 describes an international competition (BodyMaps) in collaboration with ISBI and MICCAI offering open algorithmic benchmarking AI reliability, efficiency, and generalizability in medical image segmentation. Finally, §6 concludes with a discussion of the current limitation and future promises of establishing large-scale, detailed-annotated, and multi-center datasets in medical image analysis.

2. AbdomenAtlas and Related Datasets

2.1. Preexisting Public Datasets

2.1.1. Classical Datasets (<500 CT Volumes)

Classical datasets can be categorized into two groups. *Group 1*: those designed for specific pathological conditions, such as LiTS (liver tumors), KiTS (kidney tumors), and Pancreas-CT (pancreatic tumors). These datasets usually provide more CT volumes (100s) but only annotate a specific type of anatomical structure and tumor. *Group 2*: those designed for general purposes, such as BTCV (12 structures), and WORD (16 structures). These datasets annotate more types of structures, but due to the annotation cost, they are usually of small size (10s). These classical datasets, reviewed in Table 1, have been invaluable public sources for training and validating state-of-the-art AI algorithms. As a significant advancement, our AbdomenAtlas offers 50× more CT volumes and 5× more anatomical structures (classes) than these classical datasets.

2.1.2. Abdominal Multi-Organ Segmentation (AMOS)

The AMOS dataset (Ji et al., 2022) includes 500 CT volumes and 100 MRI scans from patients with various abdominal conditions and different CT scanners. It provides detailed annotations of 15 anatomical structures and is valuable for cross-modality learning. However, the data is from only two hospitals in Asia, and important structures like the intestine and colon are not annotated. In contrast, AbdomenAtlas is much more extensive, featuring CT volumes from 47× more hospitals across 19× more countries and includes annotations for 10 additional abdominal structures.

2.1.3. AbdomenCT-1K

The AbdomenCT-1K dataset (Ma et al., 2021) provides 1,112 CT volumes from 12 hospitals, integrating data from five existing datasets and newly acquired CT volumes. It includes multi-phase, multi-vendor, and multi-disease cases. However, it only

*Correspondence to: Zongwei Zhou (zzhou82@jh.edu)
[1]Common crawl repository: https://commoncrawl.org/

Table 1. Our AbdomenAtlas consists of three component datasets—AbdomenAtlas 1.1, FullBodyAtlas-1K, and AbdomenAtlas-9K, providing a total of 20,460 annotated 3D computed tomography (CT) volumes, with many more to follow from a variety of sources. In AbdomenAtlas (exclude JHH), we employed an efficient semi-automatic annotation procedure, described in §3.3, to annotate 25 anatomical structures for 15,214 CT volumes; in JHH, a team of expert radiologists provided very high-quality annotations for 22 anatomical structures for 5,246 CT volumes as illustrated in Figure 1–Property II. AbdomenAtlas 1.1 will be made available to the public for AI training, FullBodyAtlas-1K is already publicly available for algorithmic benchmarking, and the CT volumes and annotations in AbdomenAtlas-9K will be reserved for rigorous external validation. Moreover, we invite further collaborations to expand detailed annotations on more CT volumes using our efficient human-AI synergy. It is generally agreed in the community that large-scale, detailed-annotated, and multi-center datasets are critical for AI benchmarking—the more data we use to test the algorithms, the better we can guarantee good performance in real-world conditions (e.g., clinical settings).

AbdomenAtlas components	# of volumes (original)	# of volumes (accessible)	# of annotated structures	# of hospitals	source countries	annotators
purpose: AI training						
AbdomenAtlas 1.1 (public)	9,262	9,262	25	88	MT, IE, BR, BA, AUS, TH, TW, CA, TR, CL, ES, MA, US, DE, NL, FR, IL, CN	human & AI
CHAOS (2018) [link]	40	20	1	1	TR	human
BTCV (2015) [link]	50	47	12	1	US	human
Pancreas-CT (2015) [link]	82	42	1	1	US	human
CT-ORG (2020) [link]	140	140	6	8	DE, NL, CA, FR, IL, US	human & AI
WORD (2021) [link]	150	120	16	1	CN	human
LiTS (2019) [link]	201	130	2	7	DE, NL, CA, FR, IL	human
AMOS22 (2022) [link]	500	200	15	2	CN	human & AI
KiTS (2023) [link]	600	489	3	1	US	human
AbdomenCT-1K (2021) (2023) [link]	1,062	1,000	4	12	DE, NL, CA, FR, IL, US, CN	human & AI
MSD-CT (2021) [link]	1,420	945	9	1	US	human & AI
FLARE'23 (2022) [link]	4,100	4,100	13	30	-	human & AI
Abdominal Trauma Det (2023) [link]	4,711	4,711	0	23	CL, DE, ES, TR, AUS, TH, TW, MA, MT, CA, IE, BR, BA	-
purpose: external validation						
FullBodyAtlas-1K (public)	1,761	1,761	117	9	CH, DE	human & AI
DAP Atlas (2023) [link]	533	533	142	2	DE	AI
TotalSegmentator (2022) [link]	1,228	1,228	117	7	CH	human & AI
AbdomenAtlas-9K (private)	11,223	9,437	25	15	US, CN, DE	human & AI
Pancreas-CT (test) (2015) [link]	82	38	1	1	US	human
AMOS22 (test) (2022) [link]	500	300	0	2	CN	human & AI
AutoPET (2022) [link]	1,014	445	0	2	DE	-
MSD-CT (official test) (2021) [link]	1,420	465	0	1	US	-
YF	1,224	1,224	0	1	CN	-
CirrhosisPro (2023)	1,737	1,719	2	7	CN	human
JHH (2022)	5,246	5,246	22	1	US	human

US: United States DE: Germany NL: Netherlands CA: Canada FR: France IL: Israel IE: Ireland BR: Brazil BA: Bosnia and Herzegowina
CN: China TR: Turkey CH: Switzerland AUS: Australia TH: Thailand TW: Taiwan CL: Chile ES: Spain MA: Morocco MT: Malta

annotated four structures (liver, kidney, spleen, pancreas). On the contrary, AbdomenAtlas contains annotations for 6.25× more abdominal structures, providing more comprehensive 3D human body representations.

2.1.4. *Medical Segmentation Decathlon (MSD) CT*

The MSD-CT dataset (Antonelli et al., 2021) includes 1,420 CT volumes with nine anatomical structures annotated across six segmentation tasks—making it valuable for developing generalizable medical image segmentation algorithms. Unlike the partially annotated MSD-CT dataset, AbdomenAtlas is fully annotated. We provide approximately 34K new masks, which are 35× more than those provided in the original MSD-CT dataset, as highlighted in the Figure 1–Property III.

2.1.5. *TotalSegmentator V2*

The TotalSegmentator dataset (Wasserthal et al., 2022) includes 1,228 CT volumes, focusing on whole-body segmentation of 117 anatomical structures. Derived from the University Hospital Basel, it features diverse cases, covering different ages, pathologies, scanners, body parts, and sequences.

However, there are three main issues with TotalSegmentator: *(1) CT Volumes Quality*: The CT volumes in TotalSegmentator are of lower quality due to resizing from the original 512×512 resolution down to approximately 300×300 to ease the data transfer. This resizing process inevitably results in the loss of fine-grained information. In contrast, AbdomenAtlas provides CT volumes in their original resolution, totaling 1.8 TB, and we have made every effort to ease the data transfer, such as using Huggingface, Dropbox, Google Drive, and Baidu Wangpan. *(2) Annotation Quality*: TotalSegmentator has low-quality annotations for some classes, especially tubular structures like the intestine and colon as exemplified in Figure 2, and skeletal structures like ribs and vertebrae. While TotalSegmentator only uses semi-automatic labeling, we manually annotated 5,246 CT volumes in AbdomenAtlas before using semi-automatic methods for the remaining, representing a significantly greater effort. Moreover, the annotations in TotalSegmentator were revised by two radiologists with three and six years of experience respectively. In contrast, our team includes a more experienced group of ten radiologists with three to 15 years of experience. *(3) Label Generation Procedure*: The labels in TotalSegmentator were largely produced by a single nnU-Net retrained continually as shown in Figure 1b in Wasserthal et al. (2022). Depending solely on nnU-Net could introduce a potential label bias favoring the nnU-Net architecture. This is evidenced by findings where nnFormer, UNETR, and Swin UNETR were all outperformed by nnU-Net and models building

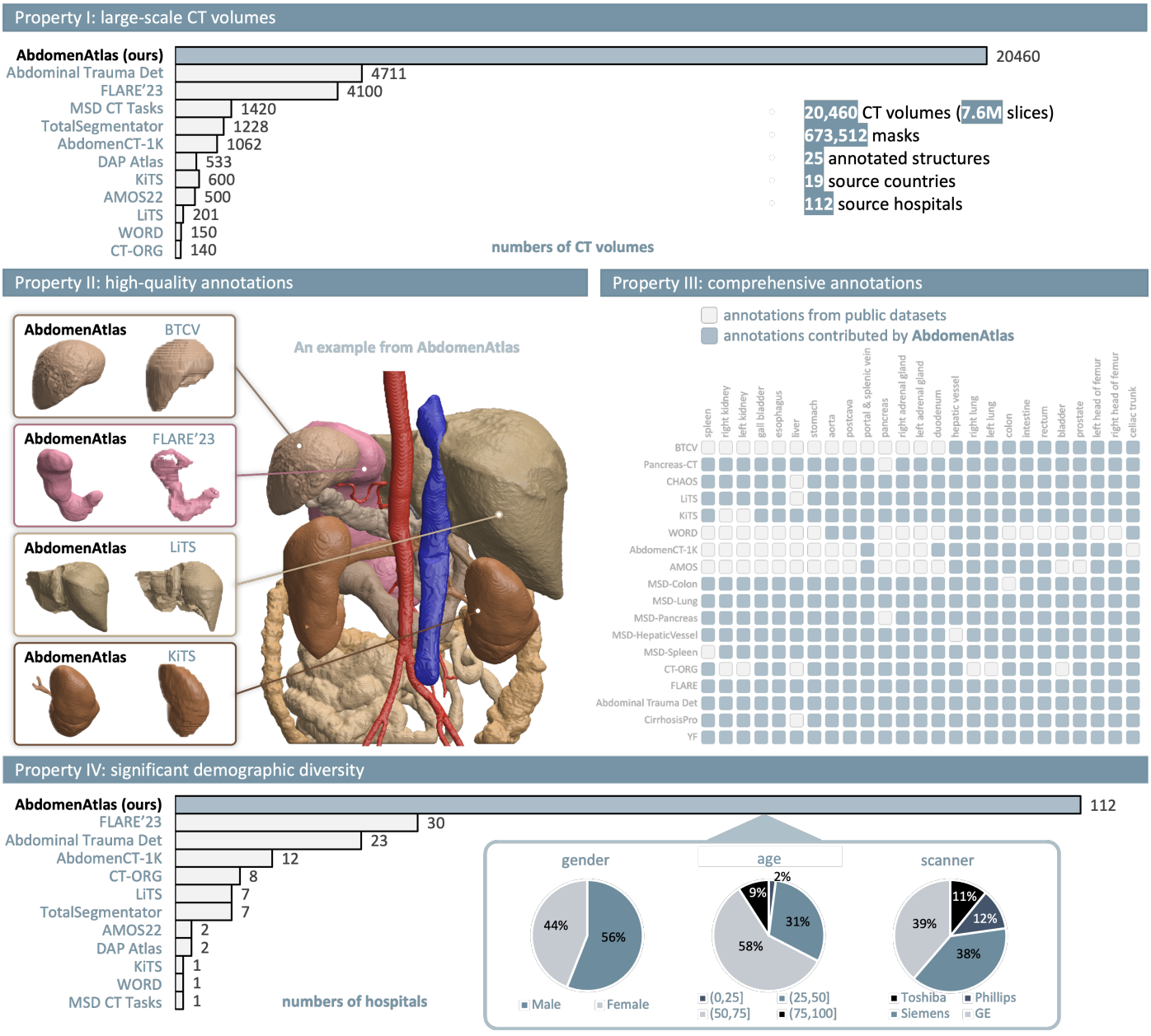


Fig. 1. *AbdomenAtlas and related datasets.* **Our AbdomenAtlas stands out from other abdominal CT datasets in four unique properties: its unprecedented scale, superior quality of annotations, comprehensive nature of these annotations, and the broad demographic diversity it includes. Property I: AbdomenAtlas provides the largest collection of annotated CT volumes among public radiology datasets, setting it apart in terms of scale. Property II: AbdomenAtlas surpasses other public radiology datasets in the quality of its annotations, offering precise boundaries and accurate representation of each anatomical structure. Property III: AbdomenAtlas provides detailed per-voxel annotations for 25 anatomical structures across each CT volume. Property IV: With CT volumes collected from 112 hospitals worldwide, AbdomenAtlas also showcases an extensive diversity in the variety of CT scanners used and patient demographics, such as sex and age.**

upon nnU-Net in TotalSegmentator (Huang et al., 2023b). To mitigate this bias, AbdomenAtlas employed three different architectures (Swin UNETR, U-Net, and nnU-net) during the semi-automatic annotation procedure.

2.1.6. *Fast, Low-resource, and Accurate oRgan and Pan-cancer sEgmentation (FLARE)*

The FLARE'23 dataset (Ma et al., 2022) contains 4,100 CT volumes from over 30 hospitals, with annotations for 13 abdominal structures and one tumor class. However, only 2,200 volumes are partially annotated, and 1,900 have no annotations. This incomplete annotation is due to the dataset's assembly from various existing datasets, each focusing on specific abdominal structures or tumors. Unlike partially annotated FLARE'23, AbdomenAtlas fully annotated 20,460 CT volumes with 25 anatomical structures. Moreover, AbdomenAtlas offers higher annotation quality as shown in the stomach example in Figure 1–Property II. Additionally, most annotated

structures in FLARE'23 are large structures that are relatively easier to detect/segment by humans and AI. AbdomenAtlas provides annotations for hard-to-segment anatomical structures, such as the hepatic vessel, intestine, and colon.

2.2. Four Properties in Our AbdomenAtlas

2.2.1. Property I: Large-Scale CT Volumes

AbdomenAtlas provides 20,460 annotated CT volumes as shown in Figure 1–Property I, associated with over 7.6M annotated CT slices. Besides providing details about AbdomenAtlas, Table 1 presents its components and subdivisions for training (AbdomenAtlas 1.1) and testing (FullBodyAtlas-1K and AbdomenAtlas-9K). AbdomenAtlasnot only represents a substantial increase in the medical data available for AI training but also serves as an extensive resource for AI benchmarking. In AbdomenAtlas, 9,262 annotated CT volumes in AbdomenAtlas 1.1 will be made available to the public for the development of AI algorithms, and 1,761 annotated CT volumes in FullBodyAtlas-1K have already been publicly available for algorithmic benchmarking, thanks to TotalSegmentator (Wasserthal et al., 2022) and DAP Atlas (Jaus et al., 2023). Moreover, we have assembled and annotated 9,437 CT volumes from 15 hospitals, termed AbdomenAtlas-9K, which will be reserved for rigorous external validation. The scale of AbdomenAtlas—20,460 CT volumes and 673K masks—allows for both the development and evaluation of AI algorithms that can apply to a wide range of medical imaging tasks.

2.2.2. Property II: High-Quality Annotations

Creating 673K high-quality masks for 25 anatomical structures requires extensive medical knowledge—at least three years of training in anatomical structures, and significant annotation costs—each structure taking about one hour for a radiologist to annotate (Park et al., 2020). As outlined in §3.1, we established a rigorous annotation standard based on human sectional anatomy (Dixon et al., 2017) to guide the radiologists in accurately annotating or revising each structure, ensuring quality control. The efficacy of this standard in maintaining our annotation quality is illustrated in Figure 1–Property II, where the annotations in our AbdomenAtlas show precise boundaries and accurate segmentation of anatomical structures, compared with those in BTCV, FLARE'23, LiTS, and KiTS.

2.2.3. Property III: Comprehensive Annotations

As depicted in Figure 1–Property III, we provide comprehensive per-voxel annotations for 25 anatomical structures, ensuring a fully-labeled dataset rather than a partially-labeled one from a naive combination of public datasets. Notably, different from the combination which only contains 39K masks, our AbdomenAtlas 1.1 provides 231K annotated structures masks for these CT volumes, substantially increasing the available masks by 5.9×. This increase not only enhances the dataset's utility but also enables the large-scale, supervised (pre-)training of AI algorithms in medical imaging analysis.

Table 2. ***Our study recruited ten radiologists, divided into two groups based on their experience.*** **The senior group consists of four radiologists with 8 to 15 years of experience in various specialties: one with 15 years in diagnostic radiology and gynecological diseases, another with 12 years in diagnostic radiology and abdominal/cerebral diseases, a third with 12 years in diagnostic radiology and thoracic diseases, and the last with 8 years in diagnostic radiology and abdominal diseases. The junior group includes six radiologists with 3 to 5 years of experience, all specializing in radiology.**

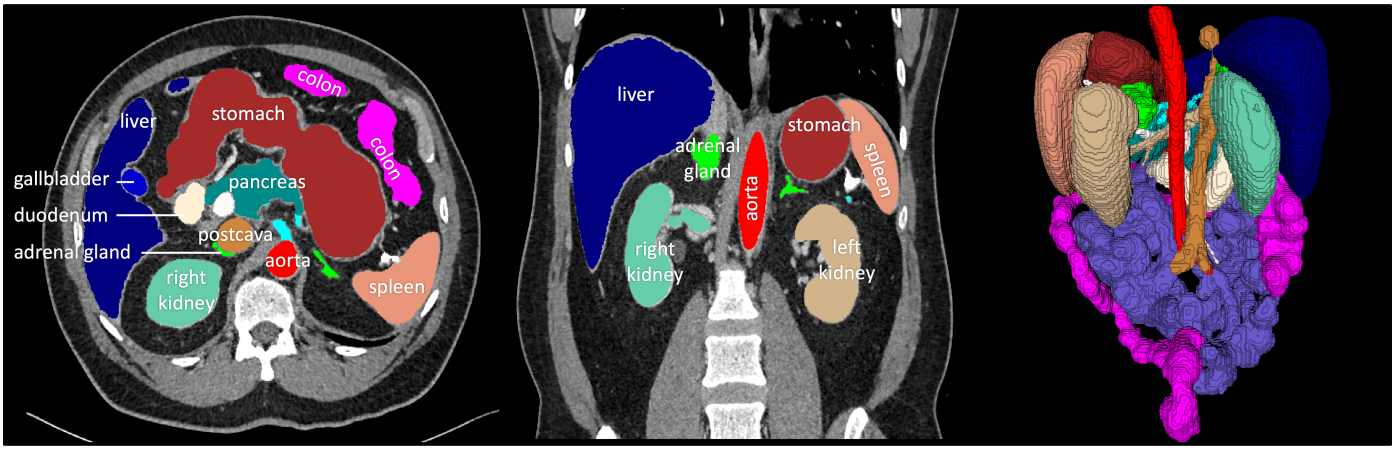
	radiologist	experience	training/expertise
senior	R1	15 years	diagnostic radiology, gynecological diseases
	R2	12 years	diagnostic radiology, abdominal/cerebral diseases
	R3	12 years	diagnostic radiology, thoracic diseases
	R4	8 years	diagnostic radiology, abdominal diseases
junior	R5	5 years	radiology
	R6	5 years	radiology
	R7	4 years	radiology
	R8	3 years	radiology
	R9	3 years	radiology
	R10	3 years	radiology

2.2.4. Property IV: Significant Demographic Diversity

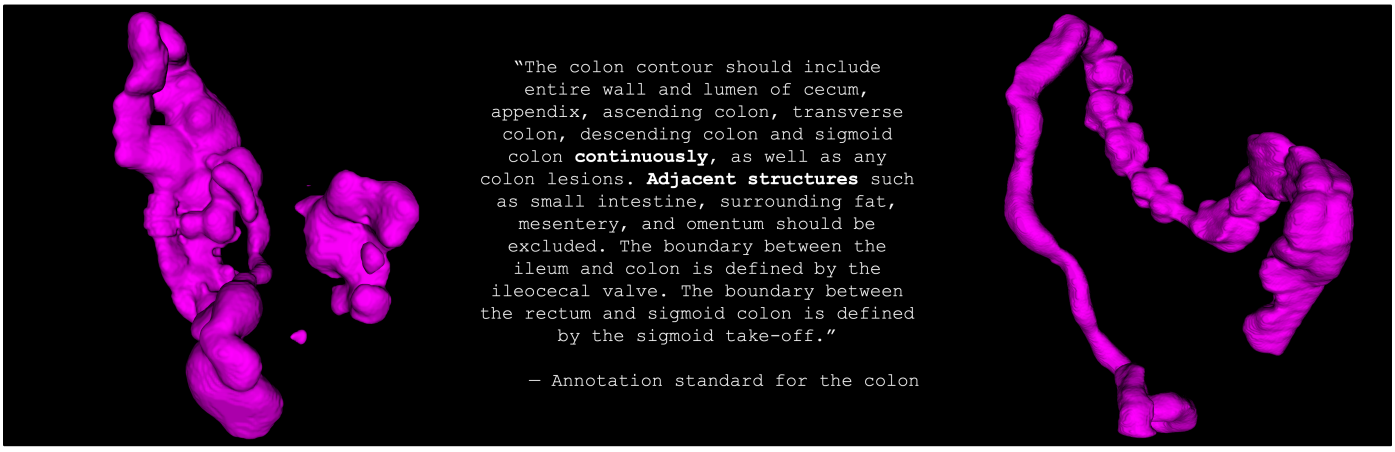
AbdomenAtlas is a multi-center dataset of pre, portal, arterial, venous, and delayed phase CT volumes collected from 112 global hospitals across eight countries. As detailed in Figure 1–Property IV, AbdomenAtlas demonstrates demographic diversity, with a balanced sex distribution of 56% female and 44% male patients and a wide age range. Notably, 58% patients aged 25 to 50 years, 31% from 50 to 75 years, 9% under 25 years, and 2% over 75 years. Moreover, AbdomenAtlas includes CT volumes from diverse scanners, such as Siemens, GE, Philips, and Toshiba, and incorporates CT volumes from both 16-/64-slice MDCT and Dual-source MDCT. These diversities in terms of phase, hospitals, countries, demography, scanners, and scan types enrich AbdomenAtlas, ensuring that AI algorithms developed with AbdomenAtlas can effectively handle variations in structure appearances influenced by different imaging protocols or patient positioning, such as rotations along the vertical axis between 30 and 60 degrees. Studies demonstrated that training data diversity is a key for AI distributional robustness (Fang et al., 2022). Accordingly, the great diversity of AbdomenAtlas contributes to developing robust, fair, and generalizable AI algorithms capable of adapting to the diverse settings found in real-world clinical environments.

3. Construction of AbdomenAtlas

Annotation quality and consistency are our top priority in constructing AbdomenAtlas. Therefore, we first established a comprehensive annotation protocol and standard (§3.1) designed to be reproducible by other teams when constructing similar datasets. Based on the standard, we applied two complementary annotation procedures. First, we adopted a manual annotation procedure—radiologists carefully annotated each CT volume voxel-by-voxel, ensuring high quality but requiring significant time investment (§3.2). Second, we adopted a semi-automatic annotation procedure combining radiologist expertise with AI algorithms (§3.3)—radiologists revised AI predictions, guided by attention maps highlighting potential errors. This human-AI synergy increased the efficiency of creating large-scale, detailed annotated datasets by 168×. Before



(a) annotation standard in AbdomenAtlas



(b) annotated colon in TotalSegmentator

(c) annotated colon in AbdomenAtlas

Fig. 2. ***Our annotation standard and annotated colon (as an example) in public datasets.*** **(a) Our datasets are annotated according to strict standards either manually or semi-automatically. These annotation standards are strictly defined, to eliminate ambiguity in physiology and anatomy, ensuring uniform and standardized annotations. In contrast, many public datasets lack comprehensive and clear standards—especially for those hard-to-segment structures like the colon—resulting in ambiguous and erroneous annotations. This discrepancy is evident when comparing (b) and (c).**

release, four senior radiologists need to verify all the annotations in AbdomenAtlas.

3.1. Annotation Protocol and Standard

Our study recruited ten radiologists, including four senior radiologists with 8 to 15 years of experience and six junior radiologists with 3 to 5 years of experience. Detailed information is presented in Table 2. For annotation accuracy and consistency, all radiologists familiarized themselves with the annotation standard as described below. We employed tools included a licensed version from Pair and an open-source 3D Slicer for annotation and revision. We provide the annotation standard for 25 structures in AbdomenAtlas, including 16 abdominal organs (esophagus, stomach, duodenum, intestine, colon, rectum, liver, gall bladder, spleen, pancreas, left kidney, right kidney, left adrenal gland, right adrenal gland, bladder, prostate), 2 thorax organs (left lung, right lung), 5 vascular structures (aorta, celiac trunk, postcava, portal & splenic vein, hepatic vessel), and 2 skeletal structures (left and right femur). An example of a detailed annotated CT volume is in Figure 2.

3.1.1. Abdominal Organs (Gastrointestinal Tract)

The stomach contour should encompass the entire stomach wall and lumen including the fundus, body, antrum, and pylorus, as well as any gastric lesions, while adjacent structures, organs, and surrounding fat should be excluded. The duodenum contour should include the entire duodenal wall and lumen from the duodenal bulb to the ligament of Treitz, along with any duodenal lesions, it should exclude surrounding structures such as the head of the pancreas, common bile duct, and surrounding vessels. The intestine contour should include the jejunum and ileum wall and lumen from the ligament of Treitz to the ileocecal valve, along with any intestinal lesions, it should exclude surrounding fat, mesentery, and mesenteric vessels. The colon contour should include the entire wall and lumen of the cecum, appendix, ascending colon, transverse colon, descending colon, and sigmoid colon, as well as any colon lesions, while adjacent structures, surrounding fat, mesentery, and omentum should be excluded. The rectum contour should include the entire rectal wall, lumen, and any lesions, while adjacent structures, surrounding fat, and muscle should be excluded.

3.1.2. Abdominal Organs (Others)

The liver contour should include all the liver parenchyma and any lesions, the intrahepatic vessels and intrahepatic bile ducts need to be covered, while excluding surrounding fat, adjacent structures, and organs. The gallbladder contour should encompass the entire gallbladder wall and lumen, including the fundus, body, and neck, as well as any gallstones or polyps, while the cystic duct, the surrounding liver parenchyma, and fat should be excluded. The pancreas contour should encompass all pancreatic parenchyma including the head, body, and tail, as well as any pancreatic lesions and pancreatic duct, the surrounding vessels and fat should be excluded. The spleen contour should include all splenic parenchyma and any lesions, it should exclude adjacent structures and extrasplenic vessels. The adrenal gland (L/R) contour should include the entire adrenal gland and any adrenal lesions, it should exclude adjacent structures and surrounding fat. The kidney (L/R) contour should include the renal parenchyma, excluding the renal pelvis, ureter, extrarenal blood vessels, surrounding fat, and any adjacent structures. The bladder contour should include the entire adrenal wall, lumen, and any bladder lesions, it should exclude adjacent structures and surrounding fat. The prostate contour should include the whole prostate parenchyma, prostatic urethra, and any prostate lesions, while excluding adjacent structures, surrounding fat, and prostatic venous plexus.

3.1.3. Thorax Organ

The esophagus contour should include the entire esophageal wall and lumen along with any esophageal lesions, while adjacent structures such as the trachea, aorta, and surrounding fat and muscle should be excluded. The lung (L/R) contour should include the entire lung parenchyma, pulmonary broncho-vascular bundle, visceral pleura, and any pulmonary lesions. It should exclude pleural effusion, pneumothorax, parietal pleura, mediastinal structures, and chest wall.

3.1.4. Vascular Structures

The aorta and celiac trunk contour should include the entire lumen of the arteries. The artery wall and calcification, ulcers, thrombosis, and dissection should also be included. The postcava and portal & splenic vein contour should include the entire lumen and cover the walls, as well as intraluminal thrombus and tumor thrombus. The hepatic vessel contour should include all intrahepatic vessel walls and lumen, as well as intraluminal thrombus and tumor thrombus.

3.1.5. Skeletal Structures

The femur (L/R) contour should include the cortical bone and spongy bone, as well as any lesions. It should exclude surrounding muscles and vessels.

3.2. Time-consuming Manual Annotation Procedure

The manual annotation procedure involves radiologists annotating each CT volume voxel by voxel according to the annotation standard defined in §3.1. While this approach can ensure accuracy and consistency, reflecting the specific needs and requirements of the data, it is time-consuming, labor-intensive, and susceptible to human errors or biases. Annotation time for a single structure may range from minutes to hours, depending on the size and complexity of the regions of interest to annotate and the local surrounding anatomical structures (Park et al., 2020). This procedure was applied to annotate the JHH dataset: A total of 22 structures for each CT volume were annotated by a team of radiologists, and confirmed by one of three additional experienced radiologists, none of whom performed the annotations, to ensure the quality of the annotation (Xia et al., 2022). JHH, involving 5,246 CT volumes, took years to complete and required the efforts of 15 radiologists.

The precision of manual annotation in AbdomenAtlas ensures that each anatomical structure is clearly defined, accurately capturing the hard-to-segment details of the body's physiology. As shown in Figure 2, AbdomenAtlas stands out in the precise manual annotation of hard-to-segment structures like the colon, demonstrating a clear advantage over datasets such as TotalSegmentator, which may have issues with annotations that erroneously include adjacent structures or are discontinuous. Despite these issues, TotalSegmentator remains a valued dataset as providing precise annotations for hard-to-segment structures is a rarity in public datasets. This is largely because manually annotating such structures is a meticulous and time-intensive task. None of the public datasets offer manual annotations for these structures across 5,246 CT volumes, setting AbdomenAtlas apart regarding both scale and annotations.

However, applying this approach to create annotations for the remaining 15,214 CT volumes in our AbdomenAtlas is extremely time-consuming. Assuming an 8-hour workday over a five-day week, a trained radiologist generally requires 60 minutes to annotate each anatomical structure within a single CT volume (Park et al., 2020). Consequently, to annotate all 15,214 CT volumes, a radiologist would need 60×25×15,214 (minutes) /60/8/5 = 9,508 (weeks) = 182.9 (years). This motivated us to develop a more efficient annotation procedure.

3.3. Semi-automatic Annotation Procedure

The semi-automatic annotation procedure combines three different AI algorithms—to minimize any bias stemming from the model architectures—on public datasets of labeled CT volumes. These AI algorithms generate initial annotations for unlabeled CT volumes. We develop an innovative strategy that can find the most important sections of the AI predictions and use color-coordinated *attention maps* to show radiologists which areas to focus on in their manual review of the AIs' work (Qu et al., 2023; Li et al., 2024). As illustrated in Figure 3, repeating this process—AI predictions and human review—over and over allowed us to accelerate the annotation process by an impressive factor of 168.

3.3.1. Attention Maps Reveals AI Mistakes

We develop attention maps to highlight potential errors in AI predictions, guiding radiologists during the review and revision process. These maps assign higher values to regions where the AI is more likely to have made mistakes, indicating areas for prioritized review. The attention map is computed using the following three criteria.

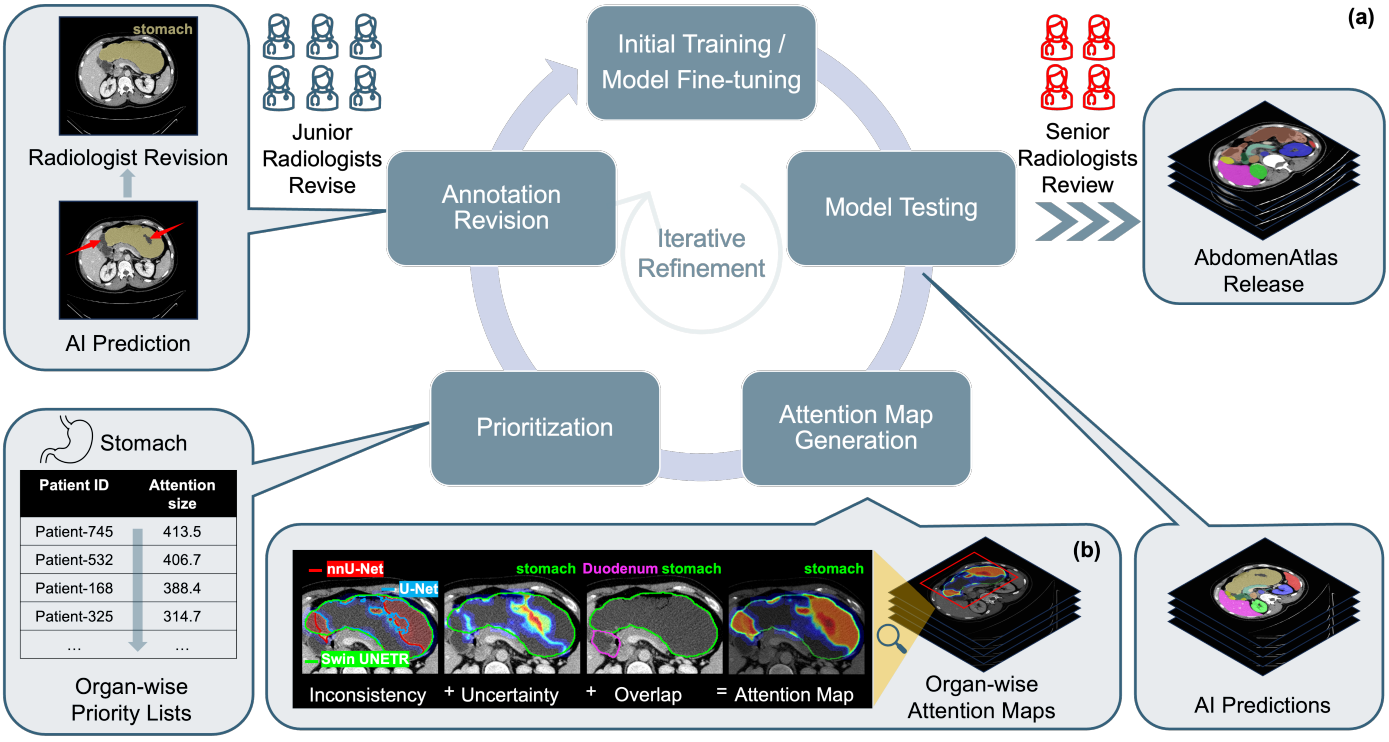


Fig. 3. ***The semi-automatic annotation procedure*** **efficiently combines the best of radiologists with AI algorithms. In this procedure, AI predictions with potential errors are manually revised by six junior radiologists under the supervision of four senior radiologists. These precise revisions play a crucial role in fine-tuning the AI algorithms, leading to a continuous improvement in their performance. Consequently, the annotations created by the AI become increasingly reliable and robust, demonstrating a successful integration of human intelligence and artificial intelligence.** ***(a) Semi-automatic annotation procedure overview.*** **In the seven-step cyclic semi-automatic annotation procedure, AI-predicted errors are highlighted using attention maps for each organ type. Subsequently, organ-wise priority lists are created, directing radiologists to focus on revising the most significant prediction errors, as signified by the largest attention map sizes. This prioritized approach ensures a more efficient and effective enhancement of the AI's predictive accuracy.** ***(b) Attention map generation.*** **The attention map integrates three criteria: inconsistency (divergence in multiple AI predictions), uncertainty (high entropy value signaling low confidence in AI predictions), and overlap (intersection of multiple organ predictions). Therefore, the attention map highlights areas with the greatest potential for prediction errors.**

1. **Inconsistency** is the standard deviation of soft predictions from three AI architectures, including Swin UNETR, nnU-Net, and U-Net. Regions with high standard deviation indicate high divergence in model predictions, prompting the need for additional manual revision.

$$\text{Inconsistency}_{i,c} = \sqrt{\frac{\sum_{n=1}^{N}(p_{i,c}^{n} - \mu_{i,c})^2}{N}}, \tag{1}$$

The subscript c denotes class c of annotated organs. For each voxel i, $p_{i,c}^{n}$ (ranging from 0 to 1) is the soft prediction value from the n-th AI architecture of class c at that voxel's index i. $\mu_{i,c}$ represents the average prediction value obtained by combining results from three AI architectures at the same voxel index. With three AI architectures (denoted by $N = 3$), $\text{Inconsistency}_{i,c}$ is determined by the standard deviation of the soft prediction values.

2. **Uncertainty**, derived from AI soft predictions' entropy values, indicates areas with reduced confidence and heightened ambiguity, potentially increasing the risk of prediction errors in those regions.

$$\text{Uncertainty}_{i,c} = -\frac{\sum_{n=1}^{N} p_{i,c}^{n} \times \log(p_{i,c}^{n})}{N}. \tag{2}$$

The $\text{Uncertainty}_{i,c}$ is averaged over different AI architectures ($N = 3$).

3. **Overlap** signifies a prediction mistake based on organ prior, specifically when a voxel is predicted to belong to multiple organs, signaling an error even in the absence of ground truth information.

$$\text{Overlap}_{i,c} = \begin{cases} 1 & \text{if } p_{i,c}^{n} > 0.5 \text{ and } \exists\, p_{i,c_{\notin}}^{n} > 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$\text{Overlap}_{i,c}$ is determined as follows: if the prediction value for class c exceeds 0.5 (a threshold value) for at least one AI architecture and there exists a prediction value not belonging to class c that exceeds 0.5 for the same voxel index i, then the overlap value is set to 1; otherwise, it is set to 0.

An attention map is created by combining inconsistency, uncertainty, and overlapping regions, facilitating annotators in efficiently identifying areas requiring revision or confirmation.

$$\text{Attention}_{i,c} = \text{Inconsistency}_{i,c} + \text{Uncertainty}_{i,c} + \text{Overlap}_{i,c} \tag{4}$$

A higher $\text{Attention}_{i,c}$ value in the attention map indicates an increased risk of prediction errors for that voxel.

3.3.2. Seven-Step Annotation Procedure

We employ a cyclical seven-step annotation procedure as illustrated in Figure 3(a), which integrates human expertise with AI predictions. This procedure is designed to enhance AI performance gradually, leading to the generation of increasingly accurate and reliable annotations.

1. **Initial Training.** An AI algorithm, denoted as $\mathcal{M}_0$ is trained from scratch utilizing 5,195 CT volumes sourced from 15 partially labeled public datasets.

2. **Model Testing.** The models[2], $\mathcal{M}_0$, undergo direct inference on 15,214 CT volumes (include AbdomenAtlas 1.1, FullBodyAtlas-1K and AbdomenAtlas-9K, excluding those from JHH) to segment 25 anatomical structures.

3. **Attention Map Generation.** For each CT volume, organ-wise attention maps are generated. These maps use a combination of inconsistency, uncertainty, and overlap metrics to identify regions potentially containing prediction errors.

4. **Prioritizing Annotation Revision.** For each anatomical structure, a priority list is generated, where AI predictions are ranked across 15,214 CT volumes based on their *attention size*, i.e. the cumulative intensity of *attention map* across all voxels. A higher *attention size* indicates an urgent requirement for manual revision due to a greater probability of prediction errors.

5. **Performing Annotation Revision.** Six junior radiologists revise the AI predictions for the top 5%[3] of samples pertaining to each anatomical structure, as identified by the organ-specific priority lists, with the guidance of the organ-specific attention maps. The remaining 95% of samples are retained without modification in Step (5). However, they will undergo automatic reassessment by the improved AI in the next cycle of the active learning process (Step 2). Samples with segmentation errors that were not selected for manual revision in an iteration of the semi-automatic annotation procedure should be selected in a future iteration, as the AI improves and other annotation errors are corrected.

6. **Model Fine-Tuning.** Using only the manually revised annotations, the current AI algorithm, $\mathcal{M}_t$, is fine-tuned to yield an improved version, $\mathcal{M}_{t+1}$.

7. **Iterative Refinement.** Steps *(2)* through *(6)* are repeated until the model's predictions for the most critical CT volumes are validated by annotators to require no further revisions, indicating minimal prediction errors.

Prior to the dataset release, four senior radiologists are responsible for a comprehensive review of the entire AbdomenAtlas and make revisions[4] if needed, particularly to address any potential errors that may remain after the semi-automatic annotation procedure is complete.

4. Application I: Efficient Transfer Learning

We present SuPreM, a suite of pre-trained 3D models that learns generalizable representations from AbdomenAtlas and provides a basis for annotation-efficient model adaptation in several applications. Specifically, SuPreM is trained for semantic segmentation on annotated CT volumes (from AbdomenAtlas 1.1) by means of supervised learning and then adapted to novel class segmentation and disease detection tasks with explicit annotations. Therefore, SuPreM have the potential to serve as Foundation Models (Bommasani et al., 2021; Moor et al., 2023). In this section, the transfer learning ability is assessed by segmentation performance on sub-datasets from FullBodyAtlas-1K and AbdomenAtlas-9K.

4.1. A Suite of Pre-trained Models: SuPreM

Our AbdomenAtlas dataset stands out in terms of its extensive scale and comprehensive annotations, offering a significant benefit for training AI models through both supervised and self-supervised manner. As of the writing of this paper, there have been no instances of either supervised or self-supervised pre-training conducted on a dataset of this magnitude (3+ million images from 9,262 volumetric data)[5]. Leveraging the comprehensive scope of our AbdomenAtlas 1.1, we have developed a suite of models (termed SuPreM). These models are built upon CNN backbones, such as U-Net and SegResNet, as well as Transformer backbones, such as Swin UNETR. As the use of pre-trained models becomes more widespread, there is an increasing need for standardized and easily accessible approaches for sharing public model weights. Accordingly, we have released a suite of pre-trained models summarized in Table 3. Releasing pre-trained foundation models should be considered a significant contribution, offering an alternative approach for knowledge sharing while simultaneously safeguarding patient privacy (Zhang and Metaxas, 2023).

To perform a fair and rigorous comparison between SuPreM and state-of-the-art supervised and self-supervised pre-trained models, we limited the SuPreM pre-training dataset to only 2,100 CT volumes from AbdomenAtlas 1.1. This size is the same as that in Liu et al. (2023b) and fewer than Tang et al. (2022) (Table 3). Table 5 shows results for training SuPreM on the entire AbdomenAtlas 1.1 and directly testing

[2]Three models (i.e., Swin UNETR, U-Net, and nnU-Net) were used in this study to reduce architectural bias.

[3]We empirically chose the threshold of 5%. First, we analyzed the initial distribution of attention size for the CT volumes in AbdomenAtlas. For most volumes, attention size was small. However, there were several significant-sized outliers, and the top 5% of CT volumes with the highest attention size captured a large proportion of these outliers. Second, manually revising 5% of the samples is feasible considering our annotation budget for each Step in the semi-automatic annotation procedure. If numerous outliers emerge or budgetary limitations exist, the threshold for revision priority should be re-calibrated.

[4]Such revisions are seldom required based on our study, with only about 100 out of 15,214 samples per anatomical structures need further adjustments.

[5]For supervised pre-training, the largest research thus far was conducted by Liu et al. (2023b), using a total of 3,410 annotated CT volumes, split into 2,100 for training and 1,310 for validation. On the other hand, the largest study in self-supervised pre-training was carried out by Tang et al. (2022), employing 5,050 CT volumes that were unannotated. **Concurrently**, Valanarasu et al. (2023) pre-trained a model on an even larger dataset, consisting of 50,000 volumes that included both CT and MRI images, using self-supervised learning.

Table 3. ***A suite of pre-trained models (SuPreM) includes several widely recognized AI models.*** **We offer pre-trained AI models, such as CNN, Transformer, and their combined versions, with plans to add more in the future. Each model was supervised pre-trained using large datasets and voxel-by-voxel annotations from AbdomenAtlas 1.1. Compared with learning from scratch and public models, fine-tuning the models in SuPreM consistently leads to the state-of-the-art performance in organ/cardiac/vertebrae/muscle/tumor segmentation on two datasets, measured by Dice Similarity Coefficient (DSC) scores. Results provided with the mean and standard deviation (mean±s.d.) are across ten trials. Additionally, we further conducted an independent two-sample *t*-test comparing the results of learning from scratch and fine-tuning models in SuPreM. The improvement in performance is statistically significant at the $P = 0.05$ level, indicated by a light blue box. Here, Tang et al. (2022), Xie et al. (2022) and Zhou et al. (2019) represent self-supervised pre-trained models, and the remaining pre-trained models employed supervised pre-training.**

pre-trained model	params	pre-trained data	TotalSegmentator organ	TotalSegmentator muscle	TotalSegmentator cardiac	TotalSegmentator vertebrae	our proprietary dataset organ	our proprietary dataset gastro	our proprietary dataset cardiac
backbone: U-Net (Ronneberger et al., 2015) and its variants									
scratch	19.08M	none	88.9±0.6	92.9±0.4	88.8±0.7	86.9±0.3	85.6±0.5	69.8±1.2	38.1±1.1
Zhou et al. (2019)	19.08M	623 CT volumes	87.8	90.1	86.3	85.1	80.1	65.5	36.9
Chen et al. (2019)	85.75M	1,638 CT volumes and masks	86.9	91.4	87.4	82.2	79.0	66.2	36.7
Xie et al. (2022)	61.79M	5,022 CT&MRI volumes	88.5	92.9	89.0	85.2	-	-	-
Zhang et al. (2021)	17.29M	920 CT volumes and masks	89.3	93.8	89.1	86.0	85.7	72.7	38.3
SuPreM	19.08M	2,100 CT volumes and masks	92.1±0.3	95.4±0.1	92.2±0.3	91.3±0.2	90.8±0.2	76.2±0.8	70.5±0.5
backbone: Swin UNETR (Hatamizadeh et al., 2021) and its variants									
scratch	62.19M	none	86.4±0.5	88.8±0.5	84.5±0.6	81.1 ±0.5	77.3±0.9	65.9± 1.7	35.5±1.4
Tang et al. (2022)	62.19M	5,050 CT volumes	89.3	93.8	88.3	86.2	87.9	72.5	38.9
Liu et al. (2023b, 2024)	62.19M	2,100 CT volumes and masks	89.7	94.1	89.4	86.5	89.1	74.6	67.6
SuPreM	62.19M	2,100 CT volumes and masks	91.3±0.3	94.6±0.2	90.3±0.3	87.2±0.3	90.4±0.7	75.9±1.2	69.8±0.9
backbone: SegResNet (Myronenko, 2019)									
scratch	4.7M	none	88.6±0.5	91.3±0.4	89.8±0.4	87.6±0.2	80.6±0.8	67.0±1.4	36.0±1.3
SuPreM	4.7M	2,100 CT volumes and masks	91.3±0.5	94.0±0.1	91.3±0.5	89.5±0.2	86.6±0.3	73.7±1.0	67.9±0.8

on AbdomenAtlas-9K. Benchmarking results indicate that, compared with learning from scratch and existing public models, models fine-tuned from SuPreM consistently achieve better segmentation performance across both TotalSegmentator and our proprietary dataset. In this section, all of the models in SuPreM follow pre-training and fine-tuning configurations as below.

- **Pre-training:** We use a random cropping method to extract sub-volumes of dimensions 96×96×96 voxels from the original CT volumes. Our SuPreM is pre-trained on AbdomenAtlas 1.1 configured with $\beta_1 = 0.9, \beta_2 = 0.999$, and a batch size of 2 per GPU, using AdamW optimizer and a cosine learning rate schedule with a warm-up for the first 100 epochs. SuPreM starts with an initial learning rate of $1e^{-4}$ and a decay of $1e^{-5}$. The pre-training has been carried out on four NVIDIA A100 using multi-GPU (4) setup with distributed data-parallel (DDP), implemented in MONAI 0.9.0, with a maximum of 800 epochs. As for the objective function of pre-training, we use the binary cross-entropy and Dice Similarity Coefficient (DSC) losses. The selection of the best model is based on achieving the highest average DSC score, calculated across 25 classes on the validation set.

- **Fine-tuning:** We fine-tune the pre-trained models using two sub-datasets from AbdomenAtlas (i.e., the TotalSegmentator from FullBodyAtlas-1K and our proprietary dataset). During fine-tuning, we maintain the initial configurations from pre-training, while modifying the warm-up schedule to 20 epochs, setting a maximum of 200 epochs, and using a single GPU. For the fine-tuning, we employ cross-entropy and DSC loss as the objective function.

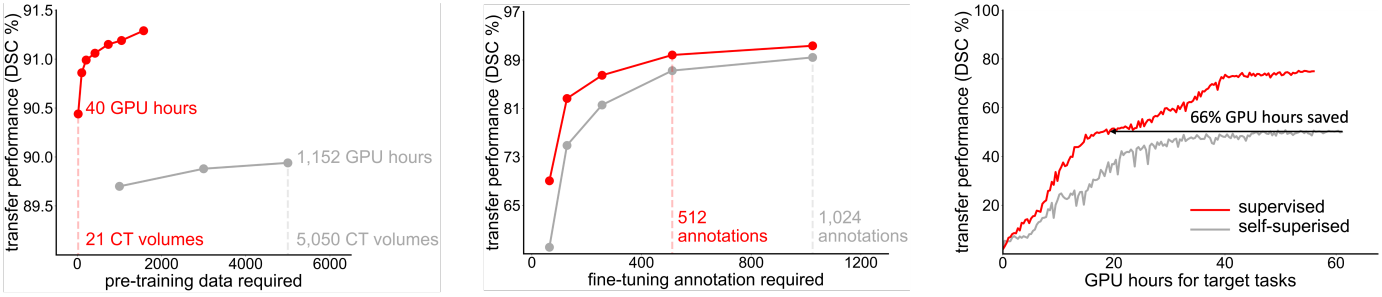
For all the other pre-training models compared in Table 3, we follow the recommended network architectures and hyper-parameter settings from their published papers for optimal performance. All the task performances are evaluated by the segmentation metric know as DSC.

4.2. *Efficiency in Transfer Learning*

In Figure 4, we present the notable efficiency of SuPreM in transferring from pretext tasks to target tasks. Specifically, (1) SuPreM pre-trained with 21 CT volumes, 672 masks, and 40 GPU hours achieves a transfer learning ability comparable to that of a self-supervised model (Tang et al., 2022) pre-trained with 5,050 CT volumes and 1,152 GPU hours. (2) SuPreM requires 50% fewer manual annotations in fine-tuning when transferring to target tasks of various anatomical structures segmentation than self-supervised pre-training. (3) SuPreM converges much faster than the self-supervised pre-trained model, resulting in 66% GPU hours saved when transferring to the target task. These results indicate the superior transfer learning efficiency of supervised pre-training with semantic segmentation.

4.2.1. *Data Efficiency*

As depicted in Figure 4(a), the need for data in SuPreM is considerably lower (21 compared to 5,050 CT volumes) than in self-supervised pre-training. This difference stems from the inherent distinct learning objectives and the information used by them. Supervised pre-training (SuPreM) gains advantages from explicit annotations, which directly guide the task, such as segmentation in this case. The model acquires knowledge from both the data and its annotations, receiving strong and precise supervision. In contrast, self-supervised learning depends on pretext tasks extracting learning features from the raw, unannotated data itself, which often results in a more unclear learning signal and necessitates a greater number of examples to capture valuable features. Notably, our results indicate that supervised pre-training scales better with increased data. When the data quantity is increased from 21 to 1,575 volumes, there is



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

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

(c) convergence & computational efficiency in *fine-tuning*

Fig. 4. ***Data, annotation, convergence, and computational efficiency.*** **For a rigorous comparison, both supervised (shown in red) and self-supervised (shown in gray) models use Swin UNETR as their backbone. The supervised pre-training is based on our SuPreM while the self-supervised pre-training uses the current state-of-the-art model (Tang et al., 2022).** ***(a)*** **The model's transfer learning ability improves consistently with an increase in the number of pre-training CT volumes, as evidenced by the results from TotalSegmentator. The model pre-trained with 21 CT volumes, 672 masks, and 40 GPU hours demonstrates a comparable transfer learning ability to that trained with 5,050 CT volumes and 1,152 GPU hours. Notably, supervised pre-training proves to be more efficient, requiring substantially 99.6% fewer data and 96.5% less computation in pre-training.** ***(b)*** **SuPreM requires 50% fewer manual annotations in fine-tuning for novel organ/cardiac/vertebrae/muscle segmentation than self-supervised pre-training models.** ***(c)*** **SuPreM reaches convergence faster than the self-supervised pre-training model, leading to a 66% GPU hours reduction in fine-tuning.**

an improvement in transfer performance on TotalSegmentator, from 90.4% to 91.3%. In comparison, in self-supervised pre-training, expanding the dataset from 1,000 to 5,050 volumes only leads to a marginal improvement in performance, from 89.7% to 89.9%. Therefore, our SuPreM not only requires substantially less data than self-supervised but also shows greater scalability and effectiveness when introducing more data.

4.2.2. *Annotation Efficiency*

We evaluated the annotation efficiency by fine-tuning SuPreM and self-supervised models (Tang et al., 2022), using varying numbers of annotated CT volumes from TotalSegmentator. As indicated in Figure 4(b), fine-tuning SuPreM can lead to a 50% reduction in manual annotation costs for anatomical structures segmentation, averaged across organs, muscles, cardiac and vertebrae which were not part of the pre-training classes. In particular, when SuPreM is fine-tuned using 512 annotated CT volumes, it shows a similar transfer learning ability to Tang et al. (2022) fine-tuned with 1,024 annotated volumes. This improvement in fine-tuning performance becomes more pronounced as the number of annotated CT volumes available for the target task is limited (e.g., 64, 128, 256).

4.2.3. *Computational Efficiency*

The computational efficiency primarily arises from the reduced data demands inherent to supervised pre-training, as previously mentioned. As depicted in Figure 4(a), SuPreM requires only 40 GPU hours to match the transfer learning performance of self-supervised pre-training, which needs 1,152 GPU hours, equating to an increase by a factor of 28.8×. In Figure 4(c), SuPreM reaches convergence more rapidly than the self-supervised pre-training for target tasks like fine-tuning on a 10% subset of TotalSegmentator, decreasing the needed GPU hours from 60 to 20. This suggests that the image features learned through supervised pre-training are inherently more representative, allowing the model to effortlessly transfer to various segmentation tasks using minimal annotated data for fine-tuning. Such computational efficiency makes supervised pre-training an attractive option for target segmentation tasks, with robust model performance, particularly when an annotated dataset is available for pre-training, such as AbdomenAtlas.

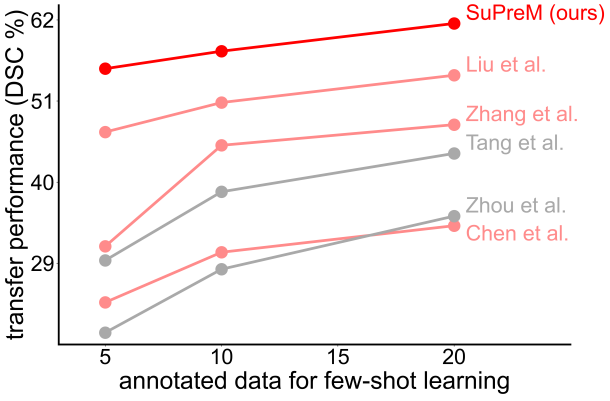


Fig. 5. We exhibit the transfer learning ability of our SuPreM on the proprietary dataset using a few examples ($N = 5, 10, 20$). Here, the transfer learning ability (represented on the Y-axis) is defined by the mean DSC score across the segmentation of 20 organ classes and 3 tumor classes. Generally, in scenarios of few-shot learning, models pre-trained with supervision (indicated in red) exhibit superior transfer ability compared to those pre-trained without supervision (shown in gray). Importantly, our SuPreM outperforms other widely recognized public models in terms of transfer learning ability.

4.3. *Transfer Learning Ability*

The transfer learning ability of SuPreM shows significant generalization and adaptability through its learned features. These features can be fine-tuned to address few-shot scenarios, segment fine-grained pancreatic tumor classes, and classify tumor sub-types with higher accuracy compared to those learned by self-supervision.

Table 4. Fine-tuning SuPreM on segmenting three novel fine-grained pancreatic tumor types from the subset of AbdomenAtlas-9K. Dice Similarity Coefficient (DSC) scores show that our SuPreM, supervised pre-trained on 2,100 annotated data, demonstrates superior transfer learning ability to three novel fine-grained pancreatic tumor classes than the self-supervised model (Tang et al., 2022) pre-trained on 5,500 raw, unlabeled data. In addition, we have further performed an independent two-sample t-test between Tang et al. (2022) and SuPreM. The performance gain (Δ) is statistically significant at the $P = 0.05$ level, with highlighting in a light blue box.

novel class	Tang et al. (2022)	SuPreM	Δ
PDAC	53.4±0.3	53.6±0.4	0.2
Cyst	41.6±0.4	49.2±0.5	7.6
PanNet	35.4±0.8	45.7±0.8	10.2
average (tumors)	48.9±0.4	53.1±0.4	4.2

4.3.1. Transfer to Few-shot Scenarios

The exploration of few-shot learning scenarios is essential to assess the adaptability and efficiency of AI models in medical image analysis, particularly when annotated data is scarce. We present the transfer performance of publicly available models and SuPreM on the proprietary dataset with a limited number (N) of annotated CT volumes (N = 5, 10, and 20). Figure 5 illustrates that SuPreM consistently outperforms other state-of-the-art models across all few-shot scenarios. The transfer performance, measured by the DSC score, shows a significant improvement as the number of annotated CT volumes increases, with SuPreM achieving the highest DSC at each few-shot scenario. This trend highlights the superior transfer learning abilities and robustness of SuPreM to deal with data scarcity.

4.3.2. Transfer to Novel Class Segmentation

The essence of transfer learning involves fine-tuning the pre-trained models to novel scenarios (Zhou et al., 2021), such as novel classes that are completely unseen during pre-training. In this study, we evaluate the transfer learning ability of SuPreM when transferred to segment novel classes that are even challenging for expert radiologists. These novel classes are three fine-grained pancreatic tumor types from a subset of AbdomenAtlas-9KȦs indicated in Table 4, when transferred to these tumor classes, our SuPreM, supervised pre-trained on anatomical structures, has better transfer learning ability than those self-supervised models pre-trained on raw, unlabeled data. We observe that the pretext task of segmentation itself inherently improves the model's ability to segment objects for novel classes. This advantage is more direct and comprehensible than the benefits offered by self-supervised tasks like contextual prediction, image masking, or instance discrimination, especially in transfer learning scenarios. We hypothesize that this is due to the supervised model learning to understand *objectness*—a concept that defines an entity as an object within an image, setting it apart from the background or other entities. Through full supervision in segmentation tasks, the model develops a deeper understanding of what characterizes an object, extending beyond the recognition of specific objects to grasp fundamental object characteristics. These characteristics include texture, boundaries, shape, size, and other vital low-level visual elements crucial for basic image segmentation.

4.3.3. Transfer to Fine-grained Tumor Identification

We have explored the transfer learning ability of SuPreM for cross-task, where transferring from anatomical structures segmentation to fine-grained tumor identification. This shift represents a significant leap, as it involves moving from segmentation tasks to classification ones, which is inherently more challenging. The difficulty in evaluating fine-grained tumor classification largely stems from the limited annotations available in public datasets, often restricted to a few hundred tumors. Our AbdomenAtlas addresses this challenge by using the subset of AbdomenAtlas-9K—containing 3,577 annotated pancreatic tumors, detailed in 1,704 PDACs, 945 Cysts, and 928 PanNets. As depicted in Figure 6(a), our findings indicate that supervised models, particularly SuPreM, demonstrate superior transfer learning abilities to classification tasks compared to self-supervised models, as evidenced by their higher Area Under the Curve (AUC) in identifying each tumor subtype. The transfer learning results, as shown in Figure 7, exhibit a sensitivity of 86.1% and specificity of 95.4% in detecting PDAC. These results exceed the average radiologist's accuracy in identifying PDAC, with a 27.6% improvement in sensitivity and 4.4% increase in specificity, according to the study (Cao et al., 2023). Additionally, SuPreM also achieved better performance in patient-level tumor detection, surpassing the sensitivity of 92.4% and specificity of 90.5% reported in the study (Xia et al., 2022), as shown by the ROC curve in Figure 6(b).

5. Application II: Open Algorithmic Benchmarking

We are organizing a medical segmentation challenge named BodyMaps, using AbdomenAtlas, hosted at the ISBI & MICCAI 2024. BodyMaps differs from many preexisting medical segmentation challenges because of the following features brought by our AbdomenAtlas. BodyMaps aims to encourage AI algorithms that are not just theoretically perform well, but also practically efficient and reliable in clinical settings.

5.1. Features of BodyMaps

5.1.1. A Large-scale Test Set

To rigorously evaluate AI algorithms, we reserved a large-scale proprietary dataset from AbdomenAtlas-9K as the test set. This dataset comprises 1,000 CT volumes with high-quality annotations, used for external validation to ensure comprehensive and reliable evaluation while reducing overfitting risk. In future editions of BodyMaps, we plan to expand this test set to 9,437 CT volumes and include 142 annotated anatomical structures, enhancing evaluation standards in medical imaging.

5.1.2. A Pronounced Domain Shift

BodyMaps focuses on AI generalizability in real-world scenarios where a pronounced domain shift often occurs when applying AI algorithms trained on one hospital's data to a different hospital. AbdomenAtlas is designed to address this problem. The test set (from AbdomenAtlas-9K private) features CT volumes with a high-resolution slice thickness of 0.5mm, significantly higher than the average 3mm resolution in the training set (from AbdomenAtlas 1.1 public). This resolution

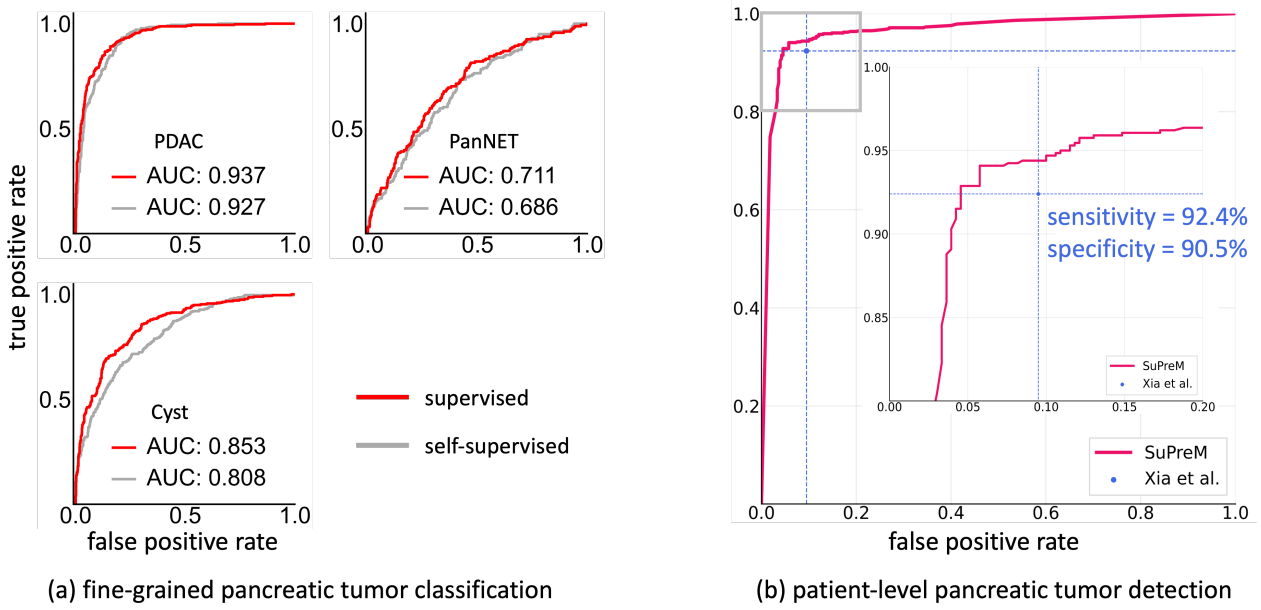


Fig. 6. ***(a) Fine-tuning SuPreM on fine-grained tumor classification involved the use of Receiver Operating Characteristic (ROC) curves to assess its transfer learning abilities in cross-task.*** **The task of identifying Cysts and PanNETs in the subset of AbdomenAtlas-9K presents unique difficulties for AI, as these lesions display a wider range of textural patterns compared to PDACs. This variation in textural patterns is evident in the Area Under the Curve (AUC) values we recorded. Across all three pancreatic tumor sub-types, the supervised pre-training model (in red) demonstrates superior transfer learning ability than the self-supervised model (Tang et al., 2022) (in gray), showing its effectiveness in fine-grained tumor classification.** ***(b) Fine-tuning SuPreM on patient-level pancreatic tumor detection task.*** **The ROC curve of SuPreM shows improved performance compared to the sensitivity of 92.4% and specificity of 90.5% as reported by Xia et al. (2022). We also provide a zoomed-in view of the ROC curve marked by a gray box to visualize the superior tumor detection performance of SuPreM.**

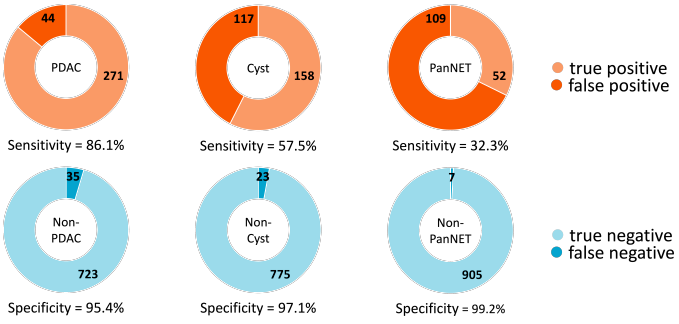


Fig. 7. ***Fine-grained pancreatic tumor classification.*** **Notably, the transfer learning results demonstrate a sensitivity of 86.1% and a specificity of 95.4% in detecting PDAC. This performance exceeds the average radiologist's performance accuracy in identifying PDAC by 27.6% in terms of sensitivity and 4.4% in specificity, as indicated in Cao et al. (2023).**

difference presents an important challenge in generalizing AI algorithms from low-resolution training sets to high-resolution test sets. In BodyMaps challenge, we reserve our test set and restrict how often participants can make their test submissions to prevent overfitting. Therefore, we recommend participants using external validation sets—such as FullBodyAtlas-1K (sourced from CH) which differs from the training set in patient demographics—to test the robustness of AI algorithms before making test submissions. Our comprehensive evaluation approach ensures that BodyMaps serves as a rigorous and fair benchmark for medical segmentation tasks, similar to the role of COCO (Lin et al., 2014) in computer vision.

Samples that are independent and identically distributed with respect to the training data are dubbed IID or in-distribution. Examples are the CT volumes in test datasets constructed by randomly splitting a database into a training and an evaluation subset. Conversely, out-of-distribution (OOD) samples are not extracted from the training data distribution. Examples are CT volumes from hospitals not contributing to the training dataset. In medical imaging applications, AI accuracy on IID test samples may be much higher than performances on OOD data (Geirhos et al., 2020). In such cases, AI may generalize poorly to real-world clinical scenarios (e.g., DeGrave et al. (2021)). We have access to test data drawn from hospitals that were never seen during training (e.g., JHH and YF in Table 1). Therefore, in BodyMaps, we evaluate generalizability across diverse clinical settings, which may encompass differences in patient demographics, equipment used, pathology prevalence and hospital protocols. Accordingly, BodyMaps fosters the creation of fair, robust and reliable AI, which presents high performance well beyond its training domain. Preliminary results of the benchmark is presented in Table 5.

5.1.3. Hard-to-segment Structures

Although advanced AI algorithms have achieved impressive performance in some organ segmentation by reaching a Dice Similarity Coefficient (DSC) score around 0.98 (e.g., liver), they still face challenges in segmenting certain anatomical structures (e.g., small objects, structures with blurry boundaries, and tubular structures like the aorta). The deficiency of AI segmenting these structures was hard to reveal using preexisting datasets, due to either the absence of such classes or the

Table 5. Preliminary results of testing AI algorithms on TotalSegmentator and JHH. We compare the out-of-distribution (OOD) performance of AI with the in-distribution (IID) performance of AI. The around 8% difference of average DSC between OOD performance and IID performance and the even larger difference for hard-to-segment structures suggests the pronounced domain shift between AbdomenAtlas 1.1 and other datasets.

class	TotalSegmentator		JHH	
	inf. SuPreM	Wasserthal et al.	inf. SuPreM	Wang et al.
spleen	95.2±0.0	98.4	95.0±0.0	97.1
kidney (R)	92.5±0.2	94.7	92.2±0.0	98.4
kidney (L)	89.0±0.3	94.4	91.6±0.1	96.8
gall bladder	82.8±0.2	84.5	83.6±0.2	90.5
liver	94.7±0.2	96.3	95.0±0.3	98.0
stomach	85.2±0.3	95.5	92.2±0.1	95.2
aorta	75.6±0.2	98.2	73.9±0.3	91.8
IVC	74.2±0.2	93.4	77.7±0.4	87.0
pancreas	83.5±0.2	89.4	79.0±0.3	87.8
average	85.9±0.2	93.9	86.7±0.2	93.6

presence of poor-quality annotations. Comparing with all pre-existing datasets, our test set provides more comprehensive and high-quality annotations of these hard-to-segment structures. This allows us to more accurately test AI algorithms, helping to identify those that perform well on hard-to-segment structures and thereby advancing the field of medical image analysis.

5.1.4. Inference Speed

We emphasize the necessity for AI algorithms to efficiently process data while maintaining high performance. Therefore, BodyMaps introduces performance metrics and ranking paradigms that consider both segmentation accuracy and inference speed, which is especially crucial in clinical environments where timely decision-making can impact patient outcomes. In addition to speed, the novel performance metrics in BodyMaps also focus on the ability of algorithms to handle complex segmentation tasks, such as those involving pronounced domain shifts and hard-to-segment structures. These complex tasks often increase computational complexity, potentially slowing down inference times. Balancing this trade-off between speed and performance is an integral challenge in BodyMaps, facilitating the development of AI solutions that are not only accurate at handling diverse scenarios but also efficient in their operation.

6. Discussion & Future Promises

6.1. Will Tumors in AbdomenAtlas be Annotated?

While the semi-automatic annotation procedure in AbdomenAtlas enables rapid scaling of annotations for various anatomical structures, it does not extend to tumor annotations. Annotating tumors is significantly more challenging due to their blurry boundaries, subtle intensity, small size, and varied conditions (Li et al., 2023; Hu et al., 2023a, 2022, 2023b). Additionally, there is no large-scale tumor dataset with detailed per-voxel annotations, making it difficult to develop strong AI algorithms for producing reasonable tumor pseudo annotations (Step 1 in our semi-automatic procedure in §3.3.2). Poor pseudo annotations would significantly increase the revision workload for radiologists. Currently, we have made some progress in annotating tumors in AbdomenAtlas. Nine types of annotated tumor examples are shown in Figure 8. We have also generated 51.8K pseudo annotations for six tumor types predicted by a combination of state-of-the-art AI algorithms (Chen et al., 2024; Lai et al., 2024). These pseudo annotations are preliminary and are pending validation by our collaborated radiologists and further verification through biopsy examinations. Based on our initial estimation, AbdomenAtlas consists of 60% normal (tumor-free) CT volumes and 40% abnormal (tumor) CT volumes, which will be invaluable resources for cancer imaging in the future.

To enhance AbdomenAtlas and address its current limitations, we are initiating a comprehensive plan focusing on the integration of tumor annotations in three possible directions. Firstly, we plan to recruit more experienced radiologists to review and revise the tumor annotations. Secondly, the integration of pathology reports as weak annotations (Xiang et al., 2023, 2024; Siddiquee et al., 2019), derived from biopsy results, will be implemented to complement the radiologists' revisions, providing a multi-faceted approach to tumor annotations that seeks to minimize annotating biases and errors. Thirdly, we intend to utilize synthetic tumor data to generate a large collection of tumor examples and their corresponding precise segmentation masks for more effective AI training and validation. These three directions are aimed at significantly enriching AbdomenAtlas with high-quality tumor annotations, thereby increasing its value for medical imaging AI development. In summary, creating annotated large-scale tumor datasets still requires significant collaborative efforts and innovative strategies from the medical imaging community.

6.2. Segment Anything vs. Our Semi-automatic Approach

An emerging research field consists of the creation of AI algorithms designed to segment arbitrary structures in medical images, according to prompts (e.g., bounding boxes) provided by the user (Ma and Wang, 2023). Such algorithms were dubbed Segment Anything Models (SAM), after their counterparts in the field of natural image segmentation (Kirillov et al., 2023). SAM can "segment anything" but does not know what is segmented. Therefore, the original SAM and its variants were limited when applied to medical images (especially for 3D volumetric data) (Ma and Wang, 2023; Huang et al., 2024; Guo et al., 2024). At the time we write this paper, SAM-based algorithms have not been integrated into standard medical annotation software (e.g., MONAI-LABEL and 3D Slicer). Once integrated, we expect SAM-based algorithms to improve the semi-automatic annotation procedure we described in this study: when revising annotations (Step 5, §3.3), radiologists can provide bounding boxes—or other weak annotations (Chou et al., 2024)—to SAM algorithms, and leverage the assistance of this interactive AI to revise an annotation more quickly (Zhang et al., 2024). Moreover, 3D promptless SAM-based algorithms (Chen et al., 2023) represent a young but promising research field. If, in the future, these algorithms start to consistently surpass standard medical segmentation AI and become the new state-of-the-art for segmenting large CT scan datasets, they could be integrated with the nnU-Net, U-Net, and Swin UNETR in the semi-automatic annotation procedure we presented in §3.3.

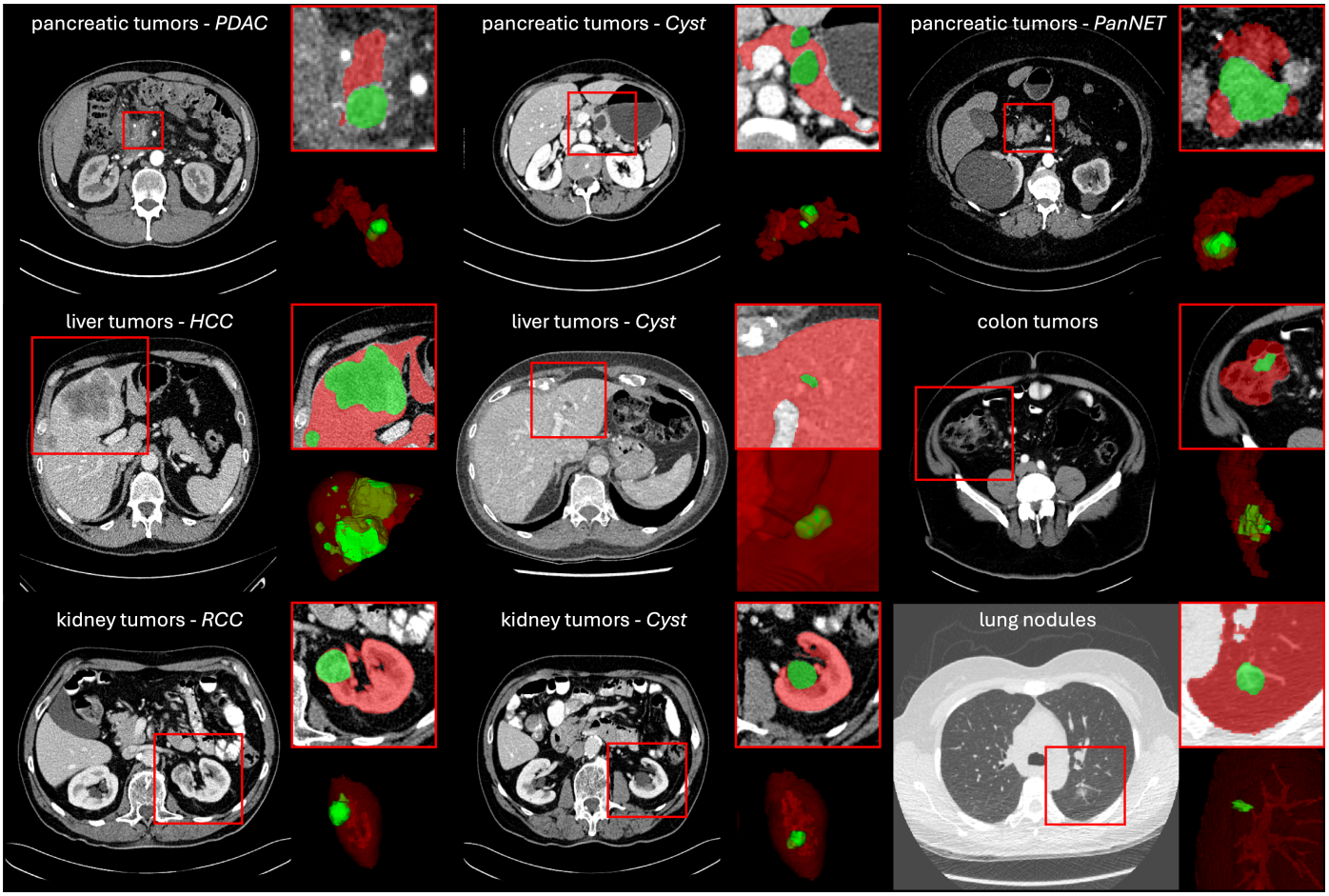


Fig. 8. ***Tumor masks in AbdomenAtlas.*** **We provide detailed per-voxel annotations of nine pathological structures in a subset of AbdomenAtlas. This includes three sub-types of pancreatic tumors: pancreatic ductal adenocarcinoma (PDAC), pancreatic neuroendocrine tumors (PanNET), and pancreatic cysts; two sub-types of liver tumors: hepatocellular carcinoma (HCC) and cysts; two sub-types of kidney tumors: renal cell carcinoma (RCC) and cysts; as well as annotations for colon tumors and lung nodules.**

6.3. Are Labels Biased to Specific AI Architectures?

The AbdomenAtlas annotations are created by a synergy between humans and AI. It is possible for annotations produced by a specific AI architecture to present patterns that are characteristic of the architecture. Thus, a U-Net may be able to better fit annotations that were originally created by another U-Net. However, three diverse architectures contribute to creating the AbdomenAtlas annotations: Swin UNETR, U-Net, and nnU-Net. This procedure increases annotation diversity and accuracy (errors characteristic of one architecture are effectively diluted by the others), and it prevents any single architecture from dominating the annotation process (by averaging the predictions from these three architectures), decreasing potential annotation bias towards a specific architecture. Additionally, a subset of AbdomenAtlas-9K was manually annotated. This characteristic further reduces the influence of potential architecture-specific bias on AI test performances reported for AbdomenAtlas.

6.4. Future Promises

We have made 9,262 annotated CT volumes from 88 hospitals publicly available for training. Moreover, a significant part of our test set is also public(FullBodyAtlas-1K, Table 1). Our private test set, AbdomenAtlas-9K, encompasses 15 hospitals and 9,437 CT volumes. Of these 9,437 CT volumes, 8,189 (86.7%) are reserved for fair evaluation which keeps both CT volumes and annotations private (JHH, CirrosisPro and YF, Table 1). The remaining 1,248 CT volumes in AbdomenAtlas-9K are public, but the annotations we created for them are currently private. These 1,248 volumes are essential for boosting the data diversity of AbdomenAtlas-9K, increasing its number of hospitals from 9 to 15. A large, diverse, and private dataset, like AbdomenAtlas-9K, is an important asset for performing third-party evaluation in algorithmic benchmarks. First, the unprecedented scale of AbdomenAtlas-9K increases the statistical significance of our BodyMaps challenge results. Second, AI performance may strongly vary across diverse hospitals (Svanera et al., 2024; Lin et al., 2024; Huang et al., 2023a). Accordingly, with 15 hospitals in AbdomenAtlas-9K, we can evaluate how well AI algorithms generalize to multiple clinical settings, and compare performances on hospitals that were seen during training, to hospitals that were never seen. Third, the private and inaccessible nature of the AbdomenAtlas-9K test

set is essential to ensure that the teams participating in our challenges cannot use this test data for training, nor overfit the test set, thus guaranteeing the integrity of the BodyMaps benchmarks' results. We will organize multiple challenges using AbdomenAtlas. After a series of challenges, we plan to publish part of AbdomenAtlas-9K progressively.

To address the current limitations of SuPreM, we plan to leverage its strengths–the extensive training on 25 anatomical structures across numerous CT volumes. This existing proficiency creates a strong foundation for enhancing SuPreM's capabilities in identifying pathological structures. Because a well-trained organ-specific foundation model can provide better accuracy and interoperability as well as significantly reduce the amount of labeled data required for new tasks, particularly tumor-related tasks, as it has already learned relevant organ features and location prior for tumors from the previous training (Zhang and Metaxas, 2023). Furthermore, we are dedicated to advancing our fine-tuning processes and integrating more advanced AI architectures and adapters to extend SuPreM's proficiency in tumor identification and adaptability to multimodal medical imaging. This strategic enhancement is designed to optimize the performance of SuPreM in addressing complex medical challenges, thereby fulfilling its promise as a transformative foundation model in medical diagnosis and treatment.

Lastly, the scope of AbdomenAtlas and AI algorithms trained on AbdomenAtlas are currently limited to a single imaging modality—CT. The medical imaging field recognizes a substantial difference in the representation of anatomical and pathological structures across different modalities, such as CT and magnetic resonance imaging (MRI). This variation presents a challenge in developing a universal dataset that is modality-comprehensive. It underscores a potential area for future expansion and highlights the necessity to extend AbdomenAtlas to a broader range of imaging modalities, thereby enhancing the robustness and applicability of AI algorithms developed using AbdomenAtlas.

Acknowledgments

This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research and the Patrick J. McGovern Foundation Award. P.R.A.S.B. thanks the funding from the Center for Biomolecular Nanotechnologies, Istituto Italiano di Tecnologia (73010, Arnesano, LE, Italy). We thank The FELIX Team at Johns Hopkins Medicine (Park et al., 2020) for collecting and annotating the JHH dataset; thank Hualin Qiao for reviewing and revising the annotation in AbdomenAtlas 1.1; thank Yu-Cheng Chou, Angtian Wang, Yaoyao Liu, Yucheng Tang, and Qi Chen for organizing the BodyMaps competition at ISBI & MICCAI-2024; thank Junfei Xiao, Jieneng Chen for their constructive suggestions at several stages of the project; thank Jaimie Patterson for disseminating our research findings. The content of this paper is covered by patents pending.

References

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., van Ginneken, B., et al., 2021. The medical segmentation decathlon. arXiv preprint arXiv:2106.05735 .

Bai, Y., Geng, X., Mangalam, K., Bar, A., Yuille, A., Darrell, T., Malik, J., Efros, A.A., 2023. Sequential modeling enables scalable learning for large vision models. arXiv preprint arXiv:2312.00785 .

Bilic, P., Christ, P.F., Vorontsov, E., Chlebus, G., Chen, H., Dou, Q., Fu, C.W., Han, X., Heng, P.A., Hesser, J., et al., 2019. The liver tumor segmentation benchmark (lits). arXiv preprint arXiv:1901.04056 .

Blankemeier, L., Cohen, J.P., Kumar, A., Van Veen, D., Gardezi, S.J.S., Paschali, M., Chen, Z., Delbrouck, J.B., Reis, E., Truyts, C., et al., 2024. Merlin: A vision language foundation model for 3d computed tomography. arXiv preprint arXiv:2406.06512 .

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al., 2021. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 .

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Advances in neural information processing systems 33, 1877–1901.

Cao, K., Xia, Y., Yao, J., Han, X., Lambert, L., Zhang, T., Tang, W., Jin, G., Jiang, H., Fang, X., et al., 2023. Large-scale pancreatic cancer detection via non-contrast ct and deep learning. Nature Medicine , 1–11.

Chen, C., Miao, J., Wu, D., Yan, Z., Kim, S., Hu, J., Zhong, A., Liu, Z., Sun, L., Li, X., Liu, T., Heng, P.A., Li, Q., 2023. Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. arXiv preprint arXiv:2309.08842 .

Chen, Q., Chen, X., Song, H., Xiong, Z., Yuille, A., Wei, C., Zhou, Z., 2024. Towards generalizable tumor synthesis, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition. URL: https://github.com/MrGiovanni/DiffTumor.

Chen, S., Ma, K., Zheng, Y., 2019. Med3d: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625 .

Chen, X., Wang, X., Zhang, K., Fung, K.M., Thai, T.C., Moore, K., Mannel, R.S., Liu, H., Zheng, B., Qiu, Y., 2022. Recent advances and clinical applications of deep learning in medical image analysis. Medical image analysis 79, 102444.

Chou, Y.C., Li, B., Fan, D.P., Yuille, A., Zhou, Z., 2024. Acquiring weak annotations for tumor localization in temporal and volumetric data. Machine Intelligence Research , 1–13URL: https://github.com/johnson111788/Drag-Drop.

Colak, E., Lin, H.M., Ball, R., Davis, M., Flanders, A., Jalal, S., Magudia, K., Marinelli, B., Nicolaou, S., Prevedello, L., Rudie, J., Shih, G., Vazirabad, M., Mongan, J., 2023. Rsna 2023 abdominal trauma detection. URL: https://kaggle.com/competitions/rsna-2023-abdominal-trauma-detection.

DeGrave, A.J., Janizek, J.D., Lee, S.I., 2021. Ai for radiographic covid-19 detection selects shortcuts over signal. Nat Mach Intell 3, 610–619. doi:10.1038/s42256-021-00338-7.

Dixon, A.K., Bowden, D.J., Logan, B.M., Ellis, H., 2017. Human sectional anatomy: Pocket atlas of body sections, CT and MRI images. CRC Press.

Fang, A., Ilharco, G., Wortsman, M., Wan, Y., Shankar, V., Dave, A., Schmidt, L., 2022. Data determines distributional robustness in contrastive language image pre-training (clip). arXiv:2205.01397.

Gatidis, S., Hepp, T., Früh, M., La Fougère, C., Nikolaou, K., Pfannenberg, C., Schölkopf, B., Küstner, T., Cyran, C., Rubin, D., 2022. A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. Scientific Data 9, 601.

Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F., 2020. Shortcut learning in deep neural networks. Nature Machine Intelligence 2, 665–673. doi:10.1038/s42256-020-00257-z.

Guo, H., Zhang, J., Huang, J., Mok, T.C.W., Guo, D., Yan, K., Lu, L., Jin, D., Xu, M., 2024. Towards a comprehensive, efficient and promptable anatomic structure segmentation model using 3d whole-body ct scans. arXiv:2403.15063.

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D., 2021. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images, in: International MICCAI Brainlesion Workshop, Springer. pp. 272–284.

Heller, N., Isensee, F., Trofimova, D., Tejpaul, R., Zhao, Z., Chen, H., Wang, L., Golts, A., Khapun, D., Shats, D., Shoshan, Y., Gilboa-Solomon, F., George, Y., Yang, X., Zhang, J., Zhang, J., Xia, Y., Wu, M., Liu, Z., Walczak, E., McSweeney, S., Vasdev, R., Hornung, C., Solaiman, R., Schoephoerster, J., Abernathy, B., Wu, D., Abdulkadir, S., Byun, B., Spriggs, J., Struyk, G., Austin, A., Simpson, B., Hagstrom, M., Virnig, S., French, J., Venkatesh, N., Chan, S., Moore, K., Jacobsen, A., Austin, S., Austin, M., Regmi, S., Papanikolopoulos, N., Weight, C., 2023. The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct. arXiv:2307.01984.

Hu, Q., Chen, Y., Xiao, J., Sun, S., Chen, J., Yuille, A.L., Zhou, Z., 2023a. Label-free liver tumor segmentation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7422–7432. URL: https://github.com/MrGiovanni/SyntheticTumors.

Hu, Q., Xiao, J., Chen, Y., Sun, S., Chen, J.N., Yuille, A., Zhou, Z., 2022. Synthetic tumors make ai segment tumors better. NeurIPS Workshop on Medical Imaging meets NeurIPS URL: https://github.com/MrGiovanni/SyntheticTumors.

Hu, Q., Yuille, A., Zhou, Z., 2023b. Synthetic data as validation. arXiv preprint arXiv:2310.16052 URL: https://github.com/MrGiovanni/SyntheticValidation.

Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., et al., 2024. Segment anything model for medical images? Medical Image Analysis 92, 103061.

Huang, Z., Deng, Z., Ye, J., Wang, H., Su, Y., Li, T., Sun, H., Cheng, J., Chen, J., He, J., et al., 2023a. A-eval: A benchmark for cross-dataset evaluation of abdominal multi-organ segmentation. arXiv preprint arXiv:2309.03906 .

Huang, Z., Wang, H., Deng, Z., Ye, J., Su, Y., Sun, H., He, J., Gu, Y., Gu, L., Zhang, S., et al., 2023b. Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. arXiv preprint arXiv:2304.06716 .

Jaus, A., Seibold, C., Hermann, K., Walter, A., Giske, K., Haubold, J., Kleesiek, J., Stiefelhagen, R., 2023. Towards unifying anatomy segmentation: Automated generation of a full-body ct dataset via knowledge aggregation and anatomical guidelines. arXiv preprint arXiv:2307.13375 .

Ji, Y., Bai, H., Yang, J., Ge, C., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., et al., 2022. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. arXiv preprint arXiv:2206.08023 .

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al., 2023. Segment anything. arXiv preprint arXiv:2304.02643 .

Lai, Y., Chen, X., Wang, A., Yuille, A., Zhou, Z., 2024. From pixel to cancer: Cellular automata in computed tomography. arXiv preprint arXiv:2403.06459 URL: https://github.com/MrGiovanni/Pixel2Cancer.

Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A., 2015. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge, in: Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge, p. 12.

Li, B., Chou, Y.C., Sun, S., Qiao, H., Yuille, A., Zhou, Z., 2023. Early detection and localization of pancreatic cancer by label-free tumor synthesis. MICCAI Workshop on Big Task Small Data, 1001-AI URL: https://github.com/MrGiovanni/SyntheticTumors.

Li, W., Yuille, A., Zhou, Z., 2024. How well do supervised models transfer to 3d image segmentation?, in: International Conference on Learning Representations. URL: https://github.com/MrGiovanni/SuPreM.

Lin, M., Weng, N., Mikolaj, K., Bashir, Z., Svendsen, M.B.S., Tolsgaard, M., Christensen, A.N., Feragen, A., 2024. Shortcut learning in medical image segmentation. arXiv preprint arXiv:2403.06748 .

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer. pp. 740–755.

Liu, H., Li, C., Wu, Q., Lee, Y.J., 2023a. Visual instruction tuning. arXiv preprint arXiv:2304.08485 .

Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., A Landman, B., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z., 2023b. Clip-driven universal model for organ segmentation and tumor detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 21152–21164. URL: https://github.com/ljwztc/CLIP-Driven-Universal-Model.

Liu, J., Zhang, Y., Wang, K., Yavuz, M.C., Chen, X., Yuan, Y., Li, H., Yang, Y., Yuille, A., Tang, Y., et al., 2024. Universal and extensible language-vision models for organ segmentation and tumor detection from abdominal computed tomography. Medical Image Analysis , 103226URL: https://github.com/ljwztc/CLIP-Driven-Universal-Model.

Luo, X., Liao, W., Xiao, J., Song, T., Zhang, X., Li, K., Wang, G., Zhang, S., 2021. Word: Revisiting organs segmentation in the whole abdominal region. arXiv preprint arXiv:2111.02403 .

Ma, J., Wang, B., 2023. Segment anything in medical images. arXiv preprint arXiv:2304.12306 .

Ma, J., Zhang, Y., Gu, S., An, X., Wang, Z., Ge, C., Wang, C., Zhang, F., Wang, Y., Xu, Y., et al., 2022. Fast and low-gpu-memory abdomen ct organ segmentation: the flare challenge. Medical Image Analysis 82, 102616.

Ma, J., Zhang, Y., Gu, S., Ge, C., Ma, S., Young, A., Zhu, C., Meng, K., Yang, X., Huang, Z., et al., 2023. Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge. arXiv preprint arXiv:2308.05862 .

Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., et al., 2021. Abdomenct-1k: Is abdominal organ segmentation a solved problem. IEEE Transactions on Pattern Analysis and Machine Intelligence .

McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.S., Darzi, A., et al., 2020. International evaluation of an ai system for breast cancer screening. Nature 577, 89–94.

Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P., 2023. Foundation models for generalist medical artificial intelligence. Nature 616, 259–265.

Myronenko, A., 2019. 3d mri brain tumor segmentation using autoencoder regularization, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4, Springer. pp. 311–320.

Park, S., Chu, L., Fishman, E., Yuille, A., Vogelstein, B., Kinzler, K., Horton, K., Hruban, R., Zinreich, E., Fouladi, D.F., et al., 2020. Annotated normal ct data of the abdomen for deep learning: Challenges and strategies for implementation. Diagnostic and interventional imaging 101, 35–44.

Qu, C., Zhang, T., Qiao, H., Liu, J., Tang, Y., Yuille, A., Zhou, Z., 2023. Abdomenatlas-8k: Annotating 8,000 abdominal ct volumes for multi-organ segmentation in three weeks. Conference on Neural Information Processing Systems URL: https://github.com/MrGiovanni/AbdomenAtlas.

Rister, B., Yi, D., Shivakumar, K., Nobashi, T., Rubin, D.L., 2020. Ct-org, a new dataset for multiple organ segmentation in computed tomography. Scientific Data 7, 1–9.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 234–241.

Roth, H.R., Lu, L., Farag, A., Shin, H.C., Liu, J., Turkbey, E.B., Summers, R.M., 2015. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 556–564.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al., 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems 35, 25278–25294.

Siddiquee, M.M.R., Zhou, Z., Tajbakhsh, N., Feng, R., Gotway, M.B., Bengio, Y., Liang, J., 2019. Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization, in: IEEE International Conference on Computer Vision, pp. 191–200. URL: https://github.com/mahfuzmohammad/Fixed-Point-GAN.

Singh, H., Mhasawade, V., Chunara, R., 2022. Generalizability challenges of mortality risk prediction models: A retrospective analysis on a multi-center database. PLOS Digital Health 1, e0000023.

Svanera, M., Savardi, M., Signoroni, A., Benini, S., Muckli, L., 2024. Fighting the scanner effect in brain mri segmentation with a progressive level-of-detail network trained on multi-site data. Medical Image Analysis 93, 103090.

Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A., 2022. Self-supervised pre-training of swin transformers for 3d medical image analysis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20730–20740.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T.,

Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al., 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 .

Valanarasu, J.M.J., Tang, Y., Yang, D., Xu, Z., Zhao, C., Li, W., Patel, V.M., Landman, B., Xu, D., He, Y., et al., 2023. Disruptive autoencoders: Leveraging low-level features for 3d medical image pre-training. arXiv preprint arXiv:2307.16896 .

Valindria, V.V., Pawlowski, N., Rajchl, M., Lavdas, I., Aboagye, E.O., Rockall, A.G., Rueckert, D., Glocker, B., 2018. Multi-modal learning from unpaired images: Application to multi-organ segmentation in ct and mri, in: 2018 IEEE winter conference on applications of computer vision (WACV), IEEE. pp. 547–556.

Wang, Y., Zhou, Y., Shen, W., Park, S., Fishman, E.K., Yuille, A.L., 2019. Abdominal multi-organ segmentation with organ-attention networks and statistical fusion. Medical image analysis 55, 88–102.

Wasserthal, J., Meyer, M., Breit, H.C., Cyriac, J., Yang, S., Segeroth, M., 2022. Totalsegmentator: robust segmentation of 104 anatomical structures in ct images. arXiv preprint arXiv:2208.05868 .

Xia, Y., Yu, Q., Chu, L., Kawamoto, S., Park, S., Liu, F., Chen, J., Zhu, Z., Li, B., Zhou, Z., et al., 2022. The felix project: Deep networks to detect pancreatic neoplasms. medRxiv .

Xiang, T., Zhang, Y., Lu, Y., Yuille, A., Zhang, C., Cai, W., Zhou, Z., 2024. Exploiting structural consistency of chest anatomy for unsupervised anomaly detection in radiography images. IEEE Transactions on Pattern Analysis and Machine Intelligence URL: https://github.com/MrGiovanni/SimSID.

Xiang, T., Zhang, Y., Lu, Y., Yuille, A.L., Zhang, C., Cai, W., Zhou, Z., 2023. Squid: Deep feature in-painting for unsupervised anomaly detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 23890–23901. URL: https://github.com/tiangexiang/SQUID.

Xie, Y., Zhang, J., Xia, Y., Wu, Q., 2022. Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier, in: European Conference on Computer Vision, Springer. pp. 558–575.

Yu, Q., Zhou, Y., Lai, Y., Qi, X., Ju, S., 2023. A multimodal deep learning network for non-invasive prediction of the hepatic decompensation risk in compensated cirrhotic people: a multicentre cohort study (chess1701). Journal of Hepatology 78, S286–S287.

Zhang, J., Xie, Y., Xia, Y., Shen, C., 2021. Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1195–1204.

Zhang, S., Metaxas, D., 2023. On the challenges and perspectives of foundation models for medical image analysis. arXiv preprint arXiv:2306.05705 .

Zhang, T., Chen, X., Qu, C., Yuille, A., Zhou, Z., 2024. Leveraging ai predicted and expert revised annotations in interactive segmentation: Continual tuning or full training?, in: IEEE International Symposium on Biomedical Imaging, IEEE. URL: https://github.com/MrGiovanni/ContinualLearning.

Zhou, Z., Sodha, V., Pang, J., Gotway, M.B., Liang, J., 2021. Models genesis. Medical Image Analysis 67, 101840. URL: https://github.com/MrGiovanni/ModelsGenesis.

Zhou, Z., Sodha, V., Siddiquee, M.M.R., Feng, R., Tajbakhsh, N., Gotway, M.B., Liang, J., 2019. Models genesis: Generic autodidactic models for 3d medical image analysis, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 384–393. URL: https://github.com/MrGiovanni/ModelsGenesis.