Scaling Tumor Segmentation: Best Lessons from Real and Synthetic Data

Qi Chen[1,2] Xinze Zhou[1] Chen Liu[1,3] Hao Chen[4] Wenxuan Li[1] Zekun Jiang[5]
Ziyan Huang[6,7] Yuxuan Zhao[8] Dexin Yu[8] Junjun He[7] Yefeng Zheng[9] Ling Shao[2]
Alan Yuille[1] Zongwei Zhou[1,*]

[1]Johns Hopkins University
[2]UCAS-Terminus AI Lab, University of Chinese Academy of Sciences
[3]Hong Kong Polytechnic University [4]University of Cambridge
[5]Sichuan University [6]Shanghai Jiao Tong University
[7]Shanghai AI Laboratory [8]Qilu Hospital of Shandong University [9]Westlake University

Code, Model & Data: https://github.com/BodyMaps/AndomenAtlas2.0

Abstract

*AI for tumor segmentation is limited by the lack of large, voxel-wise annotated datasets, which are hard to create and require medical experts. In our proprietary JHH dataset of **3,000** annotated pancreatic tumor scans, we found that AI performance stopped improving after **1,500** scans. With synthetic data, we reached the same performance using only **500** real scans. This finding suggests that synthetic data can steepen data scaling laws, enabling more efficient model training than real data alone. Motivated by these lessons, we created **AbdomenAtlas 2.0**—a dataset of 10,134 CT scans with a total of 13,223 tumor instances per-voxel manually annotated in six organs (pancreas, liver, kidney, colon, esophagus, and uterus) and 6,511 control scans. Annotated by 23 expert radiologists, it is several orders of magnitude larger than existing public tumor datasets. While we continue expanding the dataset, the current version of AbdomenAtlas 2.0 already provides a strong foundation—based on lessons from the JHH dataset—for training AI to segment tumors in six organs. It achieves notable improvements over public datasets, with a **+7%** DSC gain on in-distribution tests and **+16%** on out-of-distribution tests.*

1. Introduction

Developing AI models for tumor segmentation is fundamentally challenged by the scarcity of large, annotated datasets—owing to the immense time and expertise required for per-voxel annotation [70, 103, 107]. Inspired by scaling laws [20, 51, 78, 85], to estimate the impact of data scale on tumor segmentation performance, we first leveraged a proprietary dataset of 3,000 pancreatic tumor scans, per-voxel annotated over five years by expert radiologists and verified by pathology reports. Our previous work [61, 96] showed that this dataset enabled AI to reach radiologist-level detection accuracy. However, as shown in Figure 1, performance gains plateaued after 1,500 scans, suggesting diminishing returns from adding more real data. Recognizing that annotating 1,500 scans is still a considerable undertaking for a single tumor type, we explored the potential of synthetic data [13, 28, 40, 58, 63, 73] to further advance this plateau. By adding synthetic tumors—three times the number of real tumors—we achieved similar or

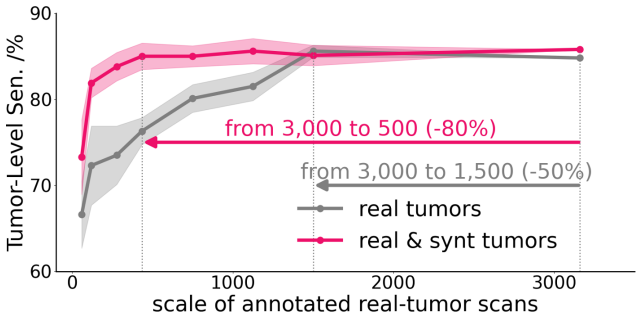


Figure 1. **Data scaling laws study.** Experimental results on the proprietary dataset demonstrate that increasing the scale of real data improve the segmentation (gray curve). Notably, supplementing the dataset with an additional 3× synthetic data (red curve) can further enhance the results, revealing the potential of a larger public dataset to advance tumor research.

*Correspondence to Zongwei Zhou (ZZHOU82@JH.EDU)

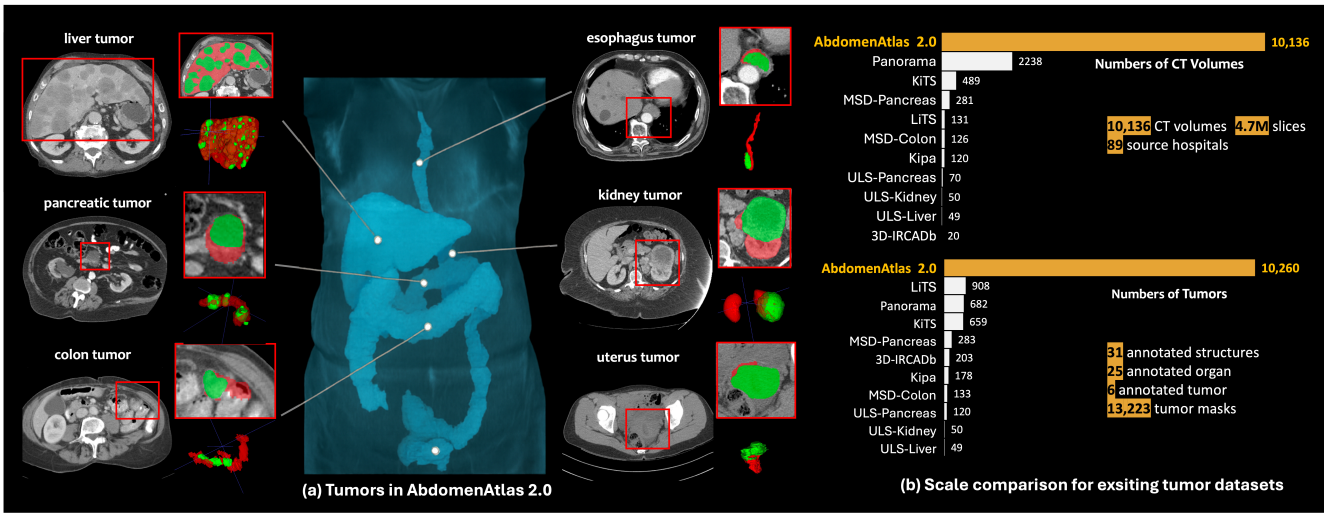


Figure 2. **Overview of the AbdomenAtlas 2.0 dataset.** For each CT scan, AbdomenAtlas 2.0 provides precise and high-quality annotations following a well-designed AI-driven annotation pipeline. Compared to existing datasets, AbdomenAtlas 2.0 collects large-scale CT scans from diverse clinical sources, encompassing a wide range of tumor types (*i.e.*, liver, pancreas, kidney, colon, esophageal, and uterine tumors) and comprehensive tumor sizes. This extensive scale makes it the largest human-annotated tumor mask dataset.

better performance with only 500 real tumor scans. This reduces annotation needs by a large margin and shows that synthetic data can accelerate learning, effectively steepening the scaling curve more than real data alone.

The lesson on the proprietary dataset helps estimate how many annotated tumor scans are needed to train effective AI models, e.g., matching radiologist performance. Considering that pancreatic tumors are especially hard to detect on CT, with 80% detected only at late stages [38, 62], we hypothesize that *if 1,500 real scans—or 500 with synthetic data—are enough for pancreatic tumors, the same or fewer might work for other organs.* Based on this idea, our first contribution is to create a dataset, that is publicly available, with 500–1,500 per-voxel annotated CT scans for tumors in six organs: pancreas, liver, kidney, colon, esophagus, and uterus. This is also the *first* public dataset that offers per-voxel annotations for esophageal and uterine tumors. We name this six-tumor dataset **AbdomenAtlas 2.0**, which comprises 3,623 CT scans with per-voxel annotations of 13,223 benign/malignant tumor instances and 6,511 normal scans as control (§3). Importantly, it includes many early-stage tumors (<20 mm): 5,709 in liver, 850 in pancreas, 3,548 in kidney, 29 in colon, 17 in esophagus, and 39 in uterus—rare and hard to collect.

While AbdomenAtlas 2.0 is much larger than public tumor datasets combined [6, 35, 49, 70], the 500–1,500 scans per tumor type are still insufficient for building robust AI across diverse data sources. This limitation is clear in our data-scaling analysis (Figure 1), where performance plateaued only on in-distribution tests. For out-of-distribution data—CT scans from different centers—performance kept improving up to 3,000 scans, suggesting that broader diversity is critical for generalization. However, scaling to that level is costly: annotating just 500–1,500 scans per tumor type required 23 radiologists and several months of effort. Selecting the most valuable scans to annotate is also challenging, since out-of-distribution data are unknown in advance.

To address this, our second contribution is to scale data and annotations through DiffTumor to produce different types of tumors (Figure 5). The data-scaling analysis (Figure 1) suggested that training AI on synthetic tumors can significantly enhances in-distribution test performance. More importantly, since collecting normal scans is much easier than acquiring and annotating tumor scans, synthetic tumors can be added to normal scans from a range of out-of-distribution sources, bypassing the need for manual per-voxel annotation. These synthetic tumors are automatically paired with per-voxel annotations as they are generated with their masks. Training AI on these normal scans augmented by synthetic tumors can greatly improve performance in out-of-distribution tests (Figure 7).

In summary, we bring data-scaling lessons from both real and synthetic data on a large proprietary dataset to develop AbdomenAtlas 2.0, achieving two key advancements for six-tumor segmentation, specifically,

1. Scaling real and synthetic data enhances performance in abdominal tumor segmentation. We rank first in the MSD challenge, leading to substantial performance improvement. We also achieve the highest performance on the validation sets of our AbdomenAtlas 2.0 dataset, improving DSC scores by +5%, +9%, +3%, +4%, +7%,

Dataset	release	# scans	# slices (K)	# tumors	tumor in #	# hospitals	# countries‡	annotators
LiTS [6] [link]	2019	131	58.6	853	liver	7	E, NL, CA, FR, IL	human
MSD-Colon [2] [link]	2021	126	13.5	131	colon	1	US	human & AI
MSD-Pancreas [2] [link]	2021	281	26.7	283	pancreas	1	US	human & AI
FLARE23 [2] [link]	2022	2,200	629.1	1,511	unknown†	30	N/A	human & AI
KiTS [36] [link]	2023	489	250.9	568	kidney	1	US	human
ULS-Liver [18] [link]	2023	49	6.3	49	liver	1	-	human
ULS-Pancreas [18] [link]	2023	120	15.4	120	pancreas	1	NI	human
ULS-Kidney [18] [link]	2023	50	6.4	50	kidney	1	N/A	human
AbdomenAtlas 2.0 (ours)	2025	10,134	4,700	13,223	liver, pancreas, kidneys, colon, esophagus, uterus	89	MT, IE, BR, BA, AUS, TH, CA, TR, CL, ES, MA, US, DE, NL, FR, IL, CN	human

† Tumors labeled in the FLARE23 dataset fall under a general 'Tumor' category without specific tumor type information.
‡ 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, CL: Chile, ES: Spain, MA: Morocco, and MT: Malta.

Table 1. **Dataset comparison.** We compare AbdomenAtlas 2.0 against existing abdominal tumor segmentation datasets, including those with and without tumor labels. AbdomenAtlas 2.0 outperforms these datasets in terms of scale and diversity.

and +2% for segmenting tumors in the liver, pancreas, kidney, colon, esophagus, and uterus, respectively, compared to the runner-up algorithms (§3.3, Tables 2–3).

2. Scaling real and synthetic data enhances generalizable performance in abdominal tumor segmentation without additional tuning and adaptation. AbdomenAtlas 2.0 significantly outperforms the runner-up algorithms by +14% DSC on four external datasets (§3.3, Table 4).

2. Related Work

Large-scale Annotated Tumor Datasets are scarce due to the limited availability of scan data and the substantial costs of obtaining per-voxel annotations. Despite these hurdles, datasets such as DeepLesion [99], AutoPET [23], PANORAMA [1], FLARE [70], and MSD [3] serve as significant efforts to mitigate this limitation. A detailed comparison of related datasets is provided in Figure 1. AbdomenAtlas 2.0 comprises more than 10,000 CT scans with voxel-level annotations across six abdominal tumors. Notably, AbdomenAtlas 2.0 features esophageal and uterine tumor scans, which have not been previously available in public datasets.

Neural Scaling Laws establish the power-law relationships that correlate model performance with key scaling factors such as model size, dataset volume, and computational resources. It is initially discovered within the domain of language models highlighted by Kaplan *et al*. [51], and soon also been observed in generative visual modeling [37, 80] and multi-modality modeling [47]. This trend of scaling underpins the recent achievements of foundation models [78, 85], emphasizing how scaling up systematically boosts model generalization and effectiveness across various tasks. However, for tumor analysis and synthetic data, scaling laws remain underexplored due to the limited availability of annotated tumor data. Leveraging our new, large-scale tumor dataset, we investigate whether similar data scaling laws exist in tumor segmentation and whether appropriate data scaling can yield a robust segmentation model capable of generalizing to detect and segment tumors from CT scans, encompassing a broad spectrum of patient demographics, imaging protocols, and healthcare facilities.

3. AbdomenAtlas 2.0

3.1. Dataset Construction

Accurate annotations are the foundation of high-quality medical datasets. However, conventional per-voxel labeling is labor-intensive. Obtaining each scan data typically costs 4–5 minutes, while extensive tumors may take up to 40 minutes [7, 70]. In addition, precisely delineating tumor boundaries takes substantial time and requires the specialized expertise of highly trained radiologists, making it impractical to scale annotations to datasets with 10,000 or more scans. To address this bottleneck, we establish a semi-automated annotation pipeline for CT scans that significantly reduces the manual workload and requires only minimal revision time from radiologists.

SMART-Annotator Procedure. Annotating missed tumors from scratch takes much longer than removing AI-generated false positives. Therefore, our annotation pipeline is designed to prioritize minimizing under-segmentation errors, thereby reducing the typical annotation time from 5 minutes per scan to less than 5 seconds on average, while maintaining high accuracy. The proposed pipeline, named SMART-Annotator, stands for *Segmentation Model-Assisted Rapid Tumor Annotator*. As depicted in Figure 3, it consists of the following four key stages:

Stage 1: Model Preparation. For each tumor, we separately train a Segmentation Model (denoted as $f(\cdot)$) using

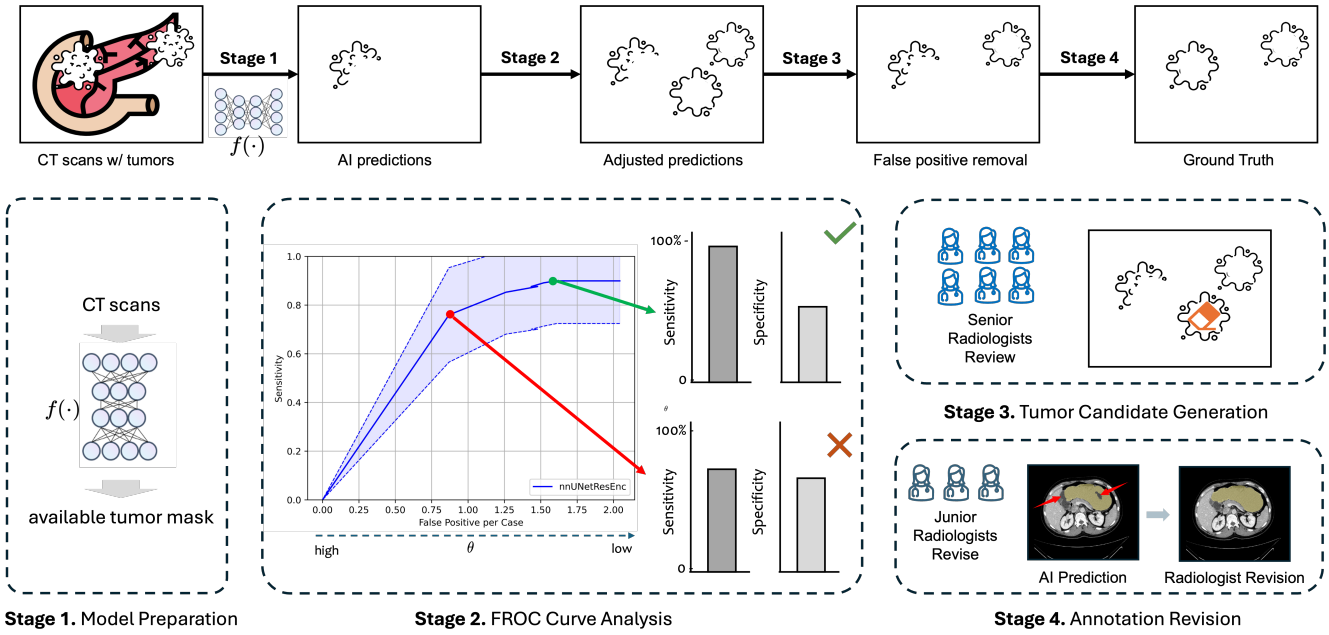


Figure 3. **Overview of the SMART-Annotator.** Towards annotating a large-scale tumor dataset, developing our SMART-Annotator involves four stages. ① Train a Segmentation Model using public datasets to provide tumor segmentation logits across AbdomenAtlas 2.0. ② Analyzing the FROC curve and selecting a threshold that enhances sensitivity to minimize missed tumors while maintaining an acceptable specificity score. ③ Removing false positives for the adjusted predictions by senior radiologists. ④ Revising the final annotations to get ground truth by junior radiologists.

publicly available datasets. The tumor-specific $f(\cdot)$ is optimized for tumor segmentation and detection tasks.

Stage 2: FROC Curve Analysis. To determine the optimal threshold, we construct the Free-response ROC (FROC) Curve by equipping $f(\cdot, \theta)$ with a set of threshold values θ, obtaining the trade-off map (as shown by the purple shadow region in Figure 3) between sensitivity and false positive rate. Experimental results on tumor analysis in CT scans reveal that a lower θ^* maximizes sensitivity while maintaining an acceptable false positive rate.

Stage 3: Tumor Candidate Generation. For CT scans requiring annotation, we apply the tumor-specific model $f(\cdot, \theta^*)$ to perform voxel-wise analysis. This process generates preliminary tumor segmentation candidates, while identifying potential tumor regions that need further refinement and validation. Since these potential regions are typically challenging, senior radiologists are then required to conduct a review to confirm true positives and eliminate false positive cases.

Stage 4: Annotation Revision. The reviewed tumor segmentation candidates undergo further refinement by junior radiologists, who annotate missed tumors and adjust mask boundaries to ensure accurate and precise tumor annotations. The final revised annotations are thoroughly reviewed by senior radiologists to guarantee high-quality ground truth.

Annotation Accuracy Analysis. For each specific organ, our pipeline adaptively adjusts the threshold θ^* based on the FROC curve to ensure over 90% sensitivity. A common concern is whether such high sensitivity might result in a significant number of false positive cases? To answer this, we validate SMART-Annotator on three public datasets and reveal that the pipeline maintains manageable false-positive rates, with an average of 1.2 false positives per scan for pancreatic tumors, 2 for liver tumors, and 2.4 for kidney tumors. These results highlight the effectiveness of our AI-driven approach in tumor detection. By pre-identifying tumors with pseudo-annotations, radiologists can quickly verify true positives, correct false positives, and, if necessary, provide additional annotations for false negatives, thereby efficiently annotating tumor scans in AbdomenAtlas 2.0.

Annotation Efficiency Analysis. The AbdomenAtlas 2.0 incorporates proprietary esophagus and uterus scans alongside unannotated data from 12 publicly available sources. Our approach applies the SMART-Annotator pipeline to all scans. Given that full manual annotation typically requires 5 minutes per scan, whereas annotation with SMART-Annotator takes only 5 seconds, this AI-driven approach substantially alleviates the annotation workload, conserving approximately $10{,}134 \times \left(5 - \frac{1}{12}\right) \approx 49,826$ minutes of valuable radiologist time for annotating the entire AbdomenAtlas 2.0 collection. Assuming a radiologist works

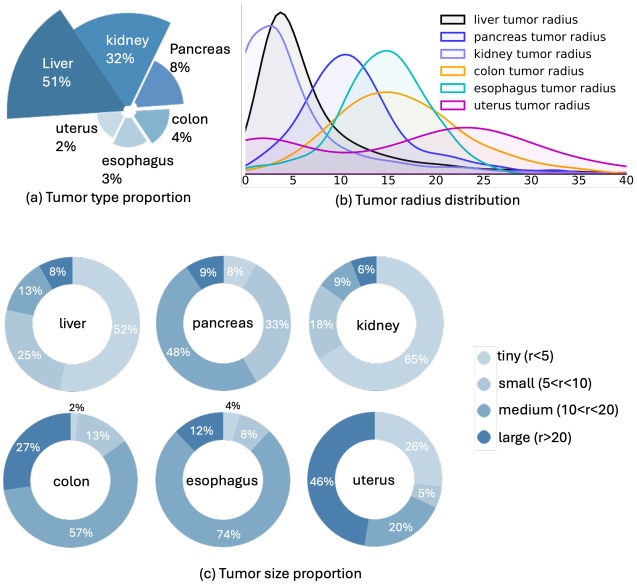


Figure 4. **Dataset statistics analysis** on the distributions of (a) different tumor proportions, (b) tumor radius, and (c) different tumor sizes categorized as tiny, small, medium, and large.

10 hours per day, this corresponds to 83 workdays saved.

3.2. Dataset Statistical Analysis

AbdomenAtlas 2.0 is the largest public, human-annotated tumor segmentation dataset, covering six tumor types (Figure 2). It improves on existing datasets in five ways:

1. Large-Scale CT Coverage. AbdomenAtlas 2.0 includes 10,136 fully annotated CT scans, totaling over 4.7 million slices. It provides labels for 31 anatomical structures, including 25 organs and 6 tumor types. The data comes from 89 hospitals, ensuring diverse patient populations and clinical conditions.

2. Diverse Tumor Types. Most public datasets focus on a single tumor type (Table 1). In contrast, AbdomenAtlas 2.0 includes liver, pancreas, kidney, colon, esophageal, and uterine tumors. It is the first public dataset with voxel-wise annotations for esophageal and uterine tumors, supporting research on rare and underrepresented cancers (Figure 4a).

3. Wide Tumor Size Range. Tumor sizes in AbdomenAtlas 2.0 range from *0* to *100* mm. We group them into four categories: tiny ($r \leq 5$ mm), small ($5 \leq r \leq 10$ mm), medium ($10 \leq r \leq 20$ mm), and large ($r \geq 20$ mm). AbdomenAtlas 2.0 provides a balanced size distribution across all tumor types, as shown in Figure 4b–c), enabling robust and scalable model training.

4. Abundant Tumor Masks. The dataset contains 10,260 annotated tumor masks across six tumor types and all size groups. It surpasses existing datasets such as LiTS and KiTS in both scale and tumor diversity (Figure 2b, Table 1).

Method	Task03 Liver		Task07 Pancreas	
	DSC	NSD	DSC	NSD
Kim *et al*. [54]	73.0	88.6	51.8	73.1
C2FNAS [102]	72.9	89.2	54.4	75.6
Trans VW [29]	76.9$_{\pm 20.0}$	92.0$_{\pm 16.8}$	51.1$_{\pm 32.8}$	70.1$_{\pm 37.4}$
Models Gen. [106]	77.5$_{\pm 20.4}$	91.9$_{\pm 17.9}$	50.4$_{\pm 32.6}$	70.0$_{\pm 37.2}$
nnU-Net [44]	76.0$_{\pm 22.1}$	90.7$_{\pm 18.3}$	52.8$_{\pm 33.0}$	71.5$_{\pm 36.6}$
DiNTS [33]	74.6$_{\pm 21.3}$	91.0$_{\pm 17.3}$	55.4$_{\pm 29.8}$	75.9$_{\pm 32.0}$
Swin UNETR [92]	75.7$_{\pm 20.4}$	91.6$_{\pm 16.8}$	58.2$_{\pm 28.6}$	79.1$_{\pm 29.7}$
Uni. Model [66]	79.4$_{\pm 17.0}$	93.4$_{\pm 15.2}$	62.3$_{\pm 26.6}$	82.9$_{\pm 27.2}$
AbdomenAtlas 2.0	**82.6**$_{\pm 11.0}$	**96.9**$_{\pm 6.4}$	**67.2**$_{\pm 24.7}$	**86.0**$_{\pm 25.2}$
Δ	**+3.2**	**+3.5**	**+4.9**	**+3.1**

Table 2. **Leaderboard performance on MSD Challenge.** The results are assessed on the MSD official server using the MSD competition test dataset. All DSC and NSD metrics are sourced from The MSD Leaderboard. The outcomes for the remaining tasks were produced by Universal Model [66, 67].

5. High-Quality Annotations. Our annotation pipeline adopts a multi-stage review process (see Figure 3), integrating AI algorithms with human expertise to enhance efficiency while maintaining high annotation quality. All images and annotations undergo rigorous quality control. This process iteratively refined the annotations until no further major revisions were necessary.

3.3. Advantages of AbdomenAtlas 2.0

Strong performance on in-distribution data. We report detailed comparisons on the official test set of the Medical Segmentation Decathlon (MSD) leaderboard in Table 2. As can be seen, with AbdomenAtlas 2.0, we significantly surpass the previously leading Universal Model [66] (denoted as Uni. Model in Table 2) and achieve the top #1 performance on the leaderboard, underscoring the superiority of AbdomenAtlas 2.0 in the task of medical segmentation.

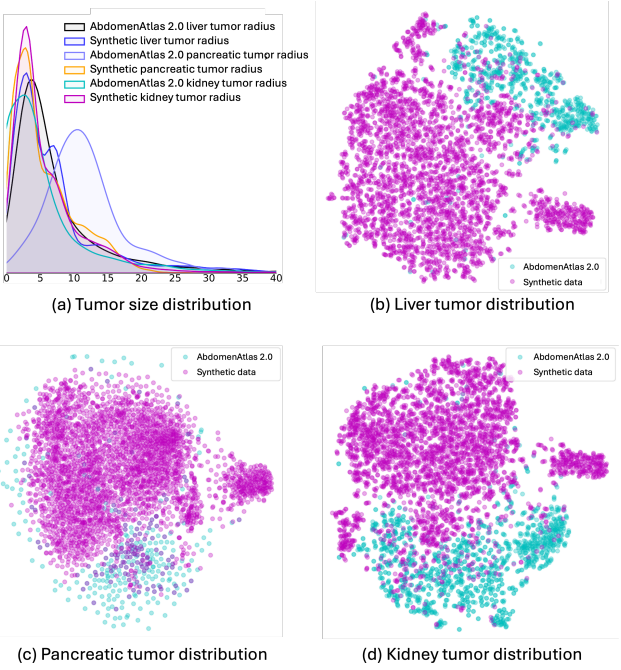
To comprehensively evaluate the six tumor types in AbdomenAtlas 2.0, we train ResEncM [45] with the annotated tumor data in AbdomenAtlas 2.0 and compare with state-of-the-art segmentation models in the medical field (*i.e*., UNETR [32], Swin UNETR [92], nnU-Net [44], ResEncM [45] and STU-Net-B [43]) that are trained with publicly available tumor datasets. The evaluations are conducted on the validation set of AbdomenAtlas 2.0 and reported in Table 3. As can be seen, training the ResEncM with AbdomenAtlas 2.0 (denoted as AbdomenAtlas 2.0) consistently improves the performance and outperforms the state-of-the-art across all tumor segmentation tasks. Compared with the second-ranked STU-Net-B, AbdomenAtlas 2.0 archives a remarkable DSC improvement of *7.3%* on esophageal tumors and *4.9%* on liver tumors, respectively. These results demonstrate the superiority of AbdomenAtlas 2.0 in delivering high-quality tumor data for model training compared to existing datasets, contributing to alleviating the data scarcity issue in tumor segmentation.

Method	Param	Liver Tumor			Pancreatic Tumor			Kidney Tumor		
		Sen.	DSC	NSD	Sen.	DSC	NSD	Sen.	DSC	NSD
UNETR [32]	101.8M	77.1 (102/131)	55.6	53.7	66.7 (102/131)	31.1	27.2	95.8 (102/131)	67.2	55.7
Swin UNETR [92]	72.8M	76.6 (102/131)	66.8	68.4	81.5 (102/131)	44.7	43.8	95.8 (102/131)	72.3	67.7
nnU-Net [44]	31.1M	80.3 (102/131)	71.7	74.6	81.5 (102/131)	56.7	54.3	100 (102/131)	84.8	80.7
ResEncM [45]	63.1M	**89.1** (102/131)	71.9	74.7	84.0 (102/131)	57.0	54.6	100 (102/131)	84.8	81.1
STU-Net-B [43]	58.3M	79.3 (102/131)	72.6	74.9	85.2 (102/131)	56.1	54.4	100 (102/131)	82.4	77.6
AbdomenAtlas 2.0	63.1M	83.7 (102/131)	**77.5**	**81.0**	**96.0** (102/131)	**65.8**	**64.7**	**100** (102/131)	**87.9**	**84.4**
Δ		**-5.4**	**+4.9**	**+6.1**	**+10.8**	**+8.8**	**+10.1**	**+0.0**	**+3.1**	**+3.3**

Method	Param	Colon Tumor			Esophagus Tumor			Uterus Tumor		
		Sen.	DSC	NSD	Sen.	DSC	NSD	Sen.	DSC	NSD
UNETR [32]	101.8M	69.2 (102/131)	27.8	29.2	92.3 (102/131)	42.3	44.1	95.8 (102/131)	69.9	60.7
Swin UNETR [92]	72.8M	65.4 (102/131)	36.8	39.4	84.6 (102/131)	48.2	49.0	95.8 (102/131)	73.8	65.0
nnU-Net [44]	31.3M	65.4 (102/131)	42.8	43.7	92.3 (102/131)	52.7	53.2	95.8 (102/131)	78.5	70.2
ResEncM [45]	63.1M	65.4 (102/131)	43.8	45.9	84.6 (102/131)	53.3	51.9	95.8 (102/131)	78.7	68.4
STU-Net-B [43]	58.3M	73.1 (102/131)	47.1	**48.7**	88.5 (102/131)	53.9	54.1	95.8 (102/131)	78.2	68.8
AbdomenAtlas 2.0	63.1M	**96.2** (102/131)	**50.7**	47.6	**96.2** (102/131)	**61.2**	**61.7**	95.8 (102/131)	**80.1**	**70.3**
Δ		**+23.1**	**+3.6**	**-1.1**	**+3.9**	**+7.3**	**+7.6**	**+0.0**	**+1.4**	**+0.1**

Table 3. **Strong performance for in-distribution data: Results on AbdomenAtlas 2.0**. We compare AbdomenAtlas 2.0 with common AI algorithms, using the validation sets from the AbdomenAtlas 2.0. AbdomenAtlas 2.0 demonstrates superior tumor segmentation and performance overall, showing significant improvements in segmenting liver tumors (+4.9%), pancreatic tumors (+8.8%), kidney tumors (+3.1%), colon tumors (+3.6%), esophagus tumors (+7.3%), and uterus tumors (+1.4%).

Better generalization for out-of-distribution data. A critical requirement for medical AI models is their ability to generalize across diverse, out-of-distribution (OOD) data from multiple hospitals, rather than being optimized solely for a single, in-distribution dataset. As shown in Table 1, AbdomenAtlas 2.0 provides a considerably more diverse collection of CT scans from 89 hospitals across 18 countries. To verify the generalizability offered by AbdomenAtlas 2.0, we further conduct evaluations on four external datasets: 3D-IRCADb [90], PANORAMA [1], Kipa [34], and a proprietary JHH dataset [96], none of which are included in the training phase. We train ResEncM [45] with the annotated tumor data in AbdomenAtlas 2.0 and compare with the following state-of-the-art medical image segmentation models: UNETR [32], Swin UNETR [92], nnU-Net [44], ResEncM [45] and STU-Net [43], SegResNet [76], Universal Model [66], and SuPrem [59]. As shown in Table 4, our model significantly outperforms previous methods on all external datasets, achieving a notable DSC improvement of *14.0%* and an NSD improvement of *17.0%* on the 3D-IRCADb dataset.



(a) Tumor size distribution (b) Liver tumor distribution

(c) Pancreatic tumor distribution (d) Kidney tumor distribution

Figure 5. **Tumor size and feature distribution of real vs. synthetic tumors.** (a) Tumor size distribution across liver, pancreatic, and kidney tumors from real and synthetic data. (b–d) Feature distributions of liver, pancreatic, and kidney tumors. We extract features using a pretrained encoder [85] and visualize them with t-SNE to compare synthetic tumors with real ones.

4. Scaling Laws in Tumor Segmentation

In this section, we explore the existence of data scaling laws in tumor segmentation and assess whether appropriate data scaling can yield a robust segmentation model. This segmentation model should be generalizable to detect and segment tumors from CT scans, handling a broad spectrum of patient demographics, imaging protocols, and healthcare facilities. Specifically, we first examine the impact of increasing the number of annotated real-tumor scans on in-distribution performance. Then we analyze how the scale of annotated real-tumor data influences the model's ability to generalize to out-of-distribution tumor data.

Method	3D-IRCADb [90] - Liver Tumor			PANORAMA [1] - Pancreatic Tumor			Kipa [34] - Kidney Tumor			JHH - Pancreatic Tumor		
	Sen.	DSC	NSD	Sen.	DSC	NSD	Sen.	DSC	NSD	Sen.	DSC	NSD
UNETR [32]	74.4 (87/117)	50.1	46.8	58.8 (77/131)	21.4	18.0	70.8 (51/72)	43.1	35.8	51.4 (152/296)	13.0	9.0
Swin UNETR [92]	76.9 (90/117)	57.9	53.7	69.5 (91/131)	34.0	30.9	81.9 (59/72)	64.3	56.6	71.3 (211/296)	31.9	21.9
nnU-Net [44]	77.8 (91/117)	65.1	62.2	75.6 (99/131)	42.4	38.6	80.6 (58/72)	64.3	58.9	69.9 (207/296)	34.1	24.7
ResEncM [45]	76.9 (90/117)	57.6	53.3	61.1 (80/131)	33.5	30.0	90.2 (65/72)	76.4	77.0	68.6 (203/296)	34.8	26.5
STU-Net [43]	78.6 (92/117)	67.1	64.5	74.0 (97/131)	42.7	40.3	55.6 (40/72)	71.2	70.4	68.9 (204/296)	34.1	24.7
SegResNet [76]	65.0 (76/117)	54.6	51.3	84.0 (110/131)	43.0	40.3	94.4 (68/72)	73.6	70.0	77.7 (211/296)	39.5	31.1
Universal Model [66]	**86.3** (101/117)	62.8	57.4	77.9 (102/131)	37.0	33.9	**97.2** (67/72)	47.8	37.1	78.4 (232/296)	32.6	27.1
SuPreM [59]	58.1 (68/117)	50.2	47.8	67.9 (89/131)	30.5	28.0	84.7 (61/72)	42.3	36.0	63.2 (187/296)	24.7	19.8
AbdomenAtlas 2.0	**86.3** (101/117)	**81.1**	**81.5**	**94.6** (124/131)	**55.3**	**52.2**	**97.2** (70/72)	**83.6**	**83.0**	**80.7** (239/296)	**45.1**	**35.7**
Δ	**+0.0**	**+14.0**	**+17.0**	**+10.6**	**+12.3**	**+11.9**	**+0.0**	**+7.2**	**+6.0**	**+2.3**	**+5.6**	**+4.6**

Table 4. **Better generalizability for out-of-distribution data: Results on external datasets.** We evaluate AbdomenAtlas 2.0 and 8 other models on data from three publicly available and one private external source without additional fine-tuning or domain adaptation. Compared to dataset-specific models, AbdomenAtlas 2.0 demonstrates greater robustness when handling CT scans obtained from a variety of scanners, protocols, and institutes.

4.1. Experimental Setup

We evaluate the scaling behavior with two data setups: (1) only real-tumor scans, and (2) a combination of both synthetic and real tumor scans. Since small tumors are rare in public datasets but crucial for clinical applications, we employ DiffTumor [12] to generate synthetic tumors, with a ratio of 4:2:1 for small, medium, and large tumors, respectively. The total number of synthetic tumor scans generated is three times of AbdomenAtlas 2.0. The distribution of tumor size and combined data distribution are illustrated in Figure 5, where we combine the generated tumors with different scales of AbdomenAtlas 2.0 training set to train the supervised ResEncM [45]. The evaluation is conducted with segmentation metrics (*i.e.*, DSC, NSD) and detection metrics (*i.e.*, tumor-level and patient level sensitivity), using the validation set of AbdomenAtlas 2.0 and six external datasets (3D-IRCADb, ULS-Liver, ULS-Pancreas, PANORAMA, Kipa, and JHH dataset).

4.2. Plateau in In-Distribution Evaluation

We report the in-distribution segmentation performance in Figure 6 and include the detection metrics in Appendix E. Our analysis of tumor segmentation scaling behavior reveals a clear trend in in-distribution performance: as the number of annotated real-tumor scans increases, the in-distribution performance gains gradually saturate. As illustrated by the gray lines in Figure 6, in-distribution performance initially improves with increasing data but eventually reaches a plateau across all three tumor types. This saturation indicates diminishing returns that adding more real tumor data yields progressively smaller performance gains.

However, combining a certain amount of synthetic tumors with real data during training helps to accelerate this in-distribution performance saturation process. As shown by the red lines in Figure 6, with the participants of synthetic data, the saturation status can be reached with only 40% to 60% of the annotated real-tumor scans, indicating

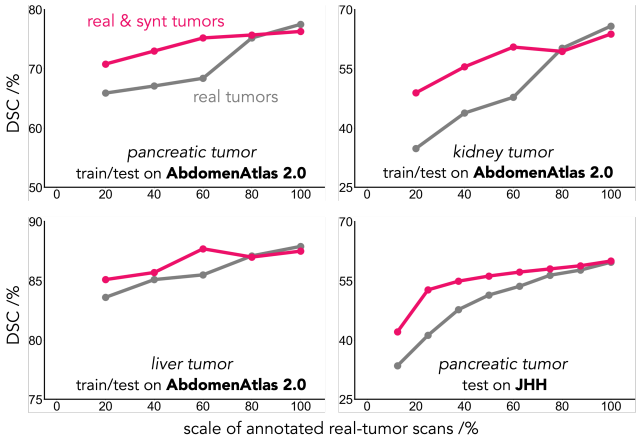


Figure 6. **Scaling data shows performance plateau in in-distribution evaluation.** We conduct a scaling study using AbdomenAtlas 2.0 and JHH datasets as real tumor data and evaluate performance on their corresponding validation sets. While scaling up the dataset initially enhances in-distribution performance, it eventually plateaus. These results align with the data-scaling lesson in §1. By supplementing real tumor data with well-designed synthetic data, we only need to collect and annotate a small amount of real data. This approach is especially beneficial for scenarios where data is scarce and annotation is costly, enabling high-accuracy segmentation with reduced effort.

that synthetic data effectively expedite the model's convergence to its optimal performance within a given domain.

This finding shows that we can achieve strong segmentation performance without collecting large amounts of real data. By supplementing real tumor data with well-designed synthetic data, we can significantly reduce the effort for costly real data annotation while maintaining strong in-distribution segmentation accuracy. This lesson demonstrates the tangible benefits of introducing synthetic data into the training process, and is particularly valuable for scenarios where real-data acquisition is costly or limited.

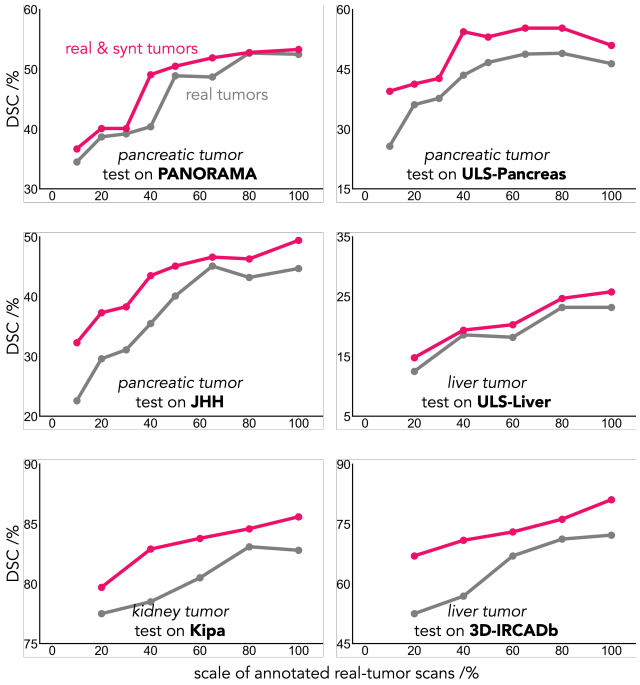


Figure 7. **Scaling data leads to greater generalizability.** We conduct a scaling study using liver, pancreatic, and kidney tumors from the Cancerverse dataset as real tumor data and evaluate performance on six external datasets. Unlike in-distribution performance, which plateaus with more data, OOD generalization continues to improve with the addition of real tumor data. Notably, the integration of synthetic data further improves generalizability, with models trained on both real and synthetic scans consistently outperforming those using only real data. These findings underscore the critical importance of data diversity in enhancing model robustness across diverse imaging conditions.

4.3. Scaling Data Leads to Greater Generalizability

Figure 7 reports out-of-distribution (OOD) segmentation performance. As indicated by the gray and red lines, OOD accuracy consistently rises with the expanding dataset. The impact of data scaling on out-of-distribution performance follows a consistently positive trend: as the amount of real tumor data increases, OOD performance continues to improve without signs of saturation, even after exhausting the entire Cancerverse dataset. We include more OOD results in Appendix F. In contrast with the in-distribution performance that tends to saturate with increasing data, the finding on the out-of-distribution performance reveals that OOD generalization continues to benefit from additional real tumor data without exhibiting diminishing returns. Such a non-diminishing trend is still obvious even when synthetic data is incorporated into the training process. Furthermore, models trained with both real and synthetic tumor scans consistently outperform those trained with real data. These findings underscore the critical role of data diversity in enhancing OOD generalization, showing that a carefully curated combination of real/synthetic data strengthens model robustness across diverse imaging settings.

5. Conclusion

This study examined not just whether more data helps tumor segmentation, but what kind of data is most valuable. Using AbdomenAtlas 2.0—a large, expert-annotated dataset—and systematic scaling experiments, we learned three key lessons: First, performance on in-distribution data plateaus early. On internal datasets like JHH, segmentation accuracy stops improving after about 1,500 real scans. This means adding more similar data yields limited benefit once a moderate threshold is reached. Second, synthetic tumors significantly reduce the need for manual annotation. By using synthetic lesions generated with DiffTumor, we can reach the same performance with only 500 real scans, cutting annotation effort by 70%. Synthetic data improves data efficiency and accelerates model convergence. Third, out-of-distribution generalization continues to improve with data diversity. Unlike the in-distribution case, performance on external datasets keeps increasing even after 1,500 scans and sees additional gains when synthetic tumors are added. This shows that model robustness depends more on data diversity than just quantity.

These lessons have important implications. Future expansion of AbdomenAtlas 2.0 should focus on including scans from different hospitals and imaging protocols. This will help the model perform better on new and unfamiliar data. For underrepresented tumor types like esophageal and uterine cancers, a few hundred well-selected scans combined with synthetic data can be enough to build useful models. AbdomenAtlas 2.0 also makes it possible to benchmark various data scaling and annotation strategies that were previously limited by small dataset size. The SMART-Annotator pipeline further shows how AI-assisted pre-labeling can reduce radiologist time from minutes to seconds per scan without sacrificing accuracy, especially when combined with synthetic tumor generation.

There are several limitations worth noting. First, the performance plateau observed at 1,500 scans applies only to pancreatic tumors in abdominal CT and with the ResEncM model. It is unclear whether this threshold holds for other organs, especially those with more complex or subtle tumor appearances. Future studies should examine how data scaling behaves across different tumor types, imaging modalities, and model architectures to see if similar saturation points occur. Second, although synthetic tumors improve performance, their anatomical realism—particularly for infiltrative, necrotic, or early-stage lesions—has not been fully verified by expert review or radiomic analysis. Ensuring clinically realistic synthesis remains a key challenge for building trust and interpretability.

Acknowledgments. This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research and the National Institutes of Health (NIH) under Award Number R01EB037669. We would like to thank the Johns Hopkins Research IT team in IT@JH for their support and infrastructure resources where some of these analyses were conducted; especially DISCOVERY HPC.

References

[1] Nuno Alves, Maarten Schuurmans, Darius Rutkowski, D. Yakar, Ingfrid Haldorsen, Marianne Liedenbaum, Anders Molven, Paolo Vendittelli, Geert Litjens, Johan Hermans, and Henk Huisman. The panorama study protocol: Pancreatic cancer diagnosis - radiologists meet ai, 2024. 3, 6, 7

[2] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, Bram van Ginneken, et al. The medical segmentation decathlon. *arXiv preprint arXiv:2106.05735*, 2021. 3, 19

[3] Michela Antonelli, Annika Reinke, Spyridon Bakas, et al. The medical segmentation decathlon. *Nature Communications*, 13:4128, 2022. 3

[4] Pedro RAS Bassi, Wenxuan Li, Jieneng Chen, Zheren Zhu, Tianyu Lin, Sergio Decherchi, Andrea Cavalli, Kang Wang, Yang Yang, Alan L Yuille, et al. Learning segmentation from radiology reports. *arXiv preprint arXiv:2507.05582*, 2025. 17

[5] Pedro RAS Bassi, Mehmet Can Yavuz, Kang Wang, Xiaoxi Chen, Wenxuan Li, Sergio Decherchi, Andrea Cavalli, Yang Yang, Alan Yuille, and Zongwei Zhou. Radgpt: Constructing 3d image-text tumor datasets. *arXiv preprint arXiv:2501.04678*, 2025. 17

[6] Patrick Bilic, Patrick Ferdinand Christ, Eugene Vorontsov, Grzegorz Chlebus, Hao Chen, Qi Dou, Chi-Wing Fu, Xiao Han, Pheng-Ann Heng, Jürgen Hesser, et al. The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*, 2019. 2, 3, 17, 19

[7] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680, 2023. 3

[8] Jingye Chen, Jieneng Chen, Zongwei Zhou, Bin Li, Alan Yuille, and Yongyi Lu. Mt-transunet: Mediating multi-task tokens in transformers for skin lesion segmentation and classification. *arXiv preprint arXiv:2112.01767*, 2021. 17

[9] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 17

[10] Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, Matthew Lungren, Lei Xing, Le Lu, Alan Yuille, and Yuyin Zhou. 3d transunet: Advancing medical image segmentation through vision transformers, 2023. 17

[11] Qi Chen, Mingxing Li, Jiacheng Li, Bo Hu, and Zhiwei Xiong. Mask rearranging data augmentation for 3d mitochondria segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 36–46. Springer, 2022. 18

[12] Qi Chen, Xiaoxi Chen, Haorui Song, Zhiwei Xiong, Alan Yuille, Chen Wei, and Zongwei Zhou. Towards generalizable tumor synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 7, 18

[13] Qi Chen, Yuxiang Lai, Xiaoxi Chen, Qixin Hu, Alan Yuille, and Zongwei Zhou. Analyzing tumors by synthesis. *arXiv preprint arXiv:2409.06035*, 2024. 1

[14] Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497, 2021. 17

[15] Douglas C Cheung and Antonio Finelli. Active surveillance in small renal masses in the elderly: a literature review. *European urology focus*, 3(4-5):340–351, 2017. 18

[16] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 424–432. Springer, 2016. 17

[17] Errol Colak, Hui-Ming Lin, Robyn Ball, Melissa Davis, Adam Flanders, Sabeena Jalal, Kirti Magudia, Brett Marinelli, Savvas Nicolaou, Luciano Prevedello, Jeff Rudie, George Shih, Maryam Vazirabad, and John Mongan. Rsna 2023 abdominal trauma detection, 2023. 19

[18] MJJ de Grauw, E Th Scholten, EJ Smit, MJCM Rutten, M Prokop, B van Ginneken, and A Hering. The uls23 challenge: a baseline model and benchmark dataset for 3d universal lesion segmentation in computed tomography. *arXiv preprint arXiv:2406.05231*, 2024. 3

[19] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012. 17

[20] Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. Scaling laws of synthetic images for model training... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7382–7392, 2024. 1

[21] Virginia Fernandez, Walter Hugo Lopez Pinaya, Pedro Borges, Petru-Daniel Tudosiu, Mark S Graham, Tom Vercauteren, and M Jorge Cardoso. Can segmentation models be trained with fully synthetically generated data? In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 79–90. Springer, 2022. 18

[22] Kathryn J Fowler, Adam Burgoyne, Tyler J Fraum, Mojgan Hosseini, Shintaro Ichikawa, Sooah Kim, Azusa Kitao, Jeong Min Lee, Valérie Paradis, Bachir Taouli, et al. Pathologic, molecular, and prognostic radiologic features of hepatocellular carcinoma. *Radiographics*, 41(6):1611–1631, 2021. 18

[23] Sergios Gatidis, Marcel Früh, Matthias P Fabritius, Sijing Gu, Konstantin Nikolaou, Christian La Fougère, Jin Ye, Junjun He, Yige Peng, Lei Bi, et al. Results from the

autopet challenge on fully automated lesion segmentation in oncologic pet/ct imaging. *Nature Machine Intelligence*, pages 1–10, 2024. 3

[24] CH Golias, A Charalabopoulos, and K Charalabopoulos. Cell proliferation and cell cycle control: a mini review. *International journal of clinical practice*, 58(12):1134–1141, 2004. 17

[25] Kuang Gong, Keith Johnson, Georges El Fakhri, Quanzheng Li, and Tinsu Pan. Pet image denoising based on denoising diffusion probabilistic model. *European Journal of Nuclear Medicine and Molecular Imaging*, pages 1–11, 2023. 18

[26] Chloe Gui, Suzanne E Kosteniuk, Jonathan C Lau, and Joseph F Megyesi. Tumor growth dynamics in serially-imaged low-grade glioma patients. *Journal of Neuro-Oncology*, 139:167–175, 2018. 18

[27] Pengfei Guo, Can Zhao, Dong Yang, Ziyue Xu, Vishwesh Nath, Yucheng Tang, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, et al. Maisi: Medical ai for synthetic imaging. *arXiv preprint arXiv:2409.11169*, 2024. 18

[28] Pengfei Guo, Can Zhao, Dong Yang, Yufan He, Vishwesh Nath, Ziyue Xu, Pedro RAS Bassi, Zongwei Zhou, Benjamin D Simon, Stephanie Anne Harmon, et al. Text2ct: Towards 3d ct volume generation from free-text descriptions using diffusion model. *arXiv preprint arXiv:2505.04522*, 2025. 1

[29] Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Zongwei Zhou, Michael B Gotway, and Jianming Liang. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. *IEEE Transactions on Medical Imaging*, 2021. 5

[30] Ibrahim Ethem Hamamci, Sezgin Er, Anjany Sekuboyina, Enis Simsar, Alperen Tezcan, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Furkan Almas, Irem Doğan, Muhammed Furkan Dasdelen, et al. Generatect: text-conditional generation of 3d chest ct volumes. In *European Conference on Computer Vision*, pages 126–143. Springer, 2025. 18

[31] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2022. 17

[32] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022. 5, 6, 7, 17

[33] Yufan He, Dong Yang, Holger Roth, Can Zhao, and Daguang Xu. Dints: Differentiable neural network topology search for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5841–5850, 2021. 5

[34] Yuting He, Guanyu Yang, Jian Yang, Rongjun Ge, Youyong Kong, Xiaomei Zhu, Shaobo Zhang, Pengfei Shao, Huazhong Shu, Jean-Louis Dillenseger, et al. Meta grayscale adaptive network for 3d integrated renal structures segmentation. *Medical image analysis*, 71:102055, 2021. 6, 7

[35] Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*, 2019. 2

[36] Nicholas Heller, Fabian Isensee, Dasha Trofimova, Resha Tejpaul, Zhongchen Zhao, Huai Chen, Lisheng Wang, Alex Golts, Daniel Khapun, Daniel Shats, et al. The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct. *arXiv preprint arXiv:2307.01984*, 2023. 3, 17, 19

[37] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020. 3

[38] Nora B Henrikson, Erin J Aiello Bowles, Paula R Blasi, Caitlin C Morrison, Matt Nguyen, Venu G Pillarisetty, and Jennifer S Lin. Screening for pancreatic cancer: updated evidence report and systematic review for the us preventive services task force. *Jama*, 322(5):445–454, 2019. 2

[39] N Hiraoka, Y Ino, S Sekine, H Tsuda, K Shimada, T Kosuge, J Zavada, M Yoshida, K Yamada, T Koyama, et al. Tumour necrosis is a postoperative prognostic marker for pancreatic cancer patients with a high interobserver reproducibility in histological evaluation. *British journal of cancer*, 103(7):1057–1065, 2010. 18

[40] Qixin Hu, Junfei Xiao, Yixiong Chen, Shuwen Sun, Jie-Neng Chen, Alan Yuille, and Zongwei Zhou. Synthetic tumors make ai segment tumors better. *NeurIPS Workshop on Medical Imaging meets NeurIPS*, 2022. 1

[41] Qixin Hu, Yixiong Chen, Junfei Xiao, Shuwen Sun, Jieneng Chen, Alan L Yuille, and Zongwei Zhou. Label-free liver tumor segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7422–7432, 2023. 18

[42] Qixin Hu, Alan Yuille, and Zongwei Zhou. Synthetic data as validation. *arXiv preprint arXiv:2310.16052*, 2023. 17

[43] Ziyan Huang, Haoyu Wang, Zhongying Deng, Jin Ye, Yanzhou Su, Hui Sun, Junjun He, Yun Gu, Lixu Gu, Shaoting Zhang, et al. Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. *arXiv preprint arXiv:2304.06716*, 2023. 5, 6, 7, 17

[44] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. 5, 6, 7

[45] Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F

Jaeger. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 488–498. Springer, 2024. 5, 6, 7

[46] Yuanfeng Ji, Haotian Bai, Jie Yang, Chongjian Ge, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *arXiv preprint arXiv:2206.08023*, 2022. 17, 19

[47] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 3

[48] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*, 2018. 17

[49] Mintong Kang, Bowen Li, Zengle Zhu, Yongyi Lu, Elliot K Fishman, Alan Yuille, and Zongwei Zhou. Label-assemble: Leveraging multiple datasets with partial labels. In *IEEE International Symposium on Biomedical Imaging*, pages 1–5. IEEE, 2023. 2

[50] Mee Joo Kang, Jin-Young Jang, Soo Jin Kim, Kyoung Bun Lee, Ji Kon Ryu, Yong-Tae Kim, Yong Bum Yoon, and Sun-Whe Kim. Cyst growth rate predicts malignancy in patients with branch duct intraductal papillary mucinous neoplasms. *Clinical Gastroenterology and Hepatology*, 9(1): 87–93, 2011. 18

[51] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 1, 3

[52] David J Kerr, Daniel G Haller, and Michael Baumann. *Oxford textbook of oncology*. oxford university press, 2016. 18

[53] Boah Kim, Yujin Oh, and Jong Chul Ye. Diffusion adversarial representation learning for self-supervised vessel segmentation. *arXiv preprint arXiv:2209.14566*, 2022. 18

[54] Sungwoong Kim, Ildoo Kim, Sungbin Lim, Woonhyuk Baek, Chiheon Kim, Hyungjoo Cho, Boogeon Yoon, and Taesup Kim. Scalable neural architecture search for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 220–228. Springer, 2019. 5

[55] Vinay Kumar, Abul Abbas, and Jon C Aster. *Robbins basic pathology*. Elsevier Health Sciences, 2017. 17, 18

[56] Yuxiang Lai, Xiaoxi Chen, Angtian Wang, Alan Yuille, and Zongwei Zhou. From pixel to cancer: Cellular automata in computed tomography. *arXiv preprint arXiv:2403.06459*, 2024. 18

[57] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, page 12, 2015. 19

[58] Bowen Li, Yu-Cheng Chou, Shuwen Sun, Hualin Qiao, Alan Yuille, and Zongwei Zhou. Early detection and localization of pancreatic cancer by label-free tumor synthesis. *MICCAI Workshop on Big Task Small Data, 1001-AI*, 2023. 1

[59] Wenxuan Li, Chongyu Qu, Xiaoxi Chen, Pedro RAS Bassi, Yijia Shi, Yuxiang Lai, Qian Yu, Huimin Xue, Yixiong Chen, Xiaorui Lin, et al. Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. *Medical Image Analysis*, page 103285, 2024. 6, 7

[60] Wenxuan Li, Alan Yuille, and Zongwei Zhou. How well do supervised models transfer to 3d image segmentation? In *International Conference on Learning Representations*, 2024. 17

[61] Wenxuan Li, Pedro RAS Bassi, Tianyu Lin, Yu-Cheng Chou, Xinze Zhou, Yucheng Tang, Fabian Isensee, Kang Wang, Qi Chen, Xiaowei Xu, et al. Scalemai: Accelerating the development of trusted datasets and ai models. *arXiv preprint arXiv:2501.03410*, 2025. 1, 17

[62] Wenxuan Li, Xinze Zhou, Qi Chen, Tianyu Lin, Pedro RAS Bassi, Szymon Plotka, Jaroslaw B Cwikla, Xiaoxi Chen, Chen Ye, Zheren Zhu, et al. Pants: The pancreatic tumor segmentation dataset. *arXiv preprint arXiv:2507.01291*, 2025. 2

[63] Xinran Li, Yi Shuai, Chen Liu, Qi Chen, Qilong Wu, Pengfei Guo, Dong Yang, Can Zhao, Pedro RAS Bassi, Daguang Xu, et al. Text-driven tumor synthesis. *arXiv preprint arXiv:2412.18589*, 2024. 1

[64] Tianyu Lin, Xinran Li, Chuntung Zhuang, Qi Chen, Yuanhao Cai, Kai Ding, Alan L Yuille, and Zongwei Zhou. Are pixel-wise metrics reliable for sparse-view computed tomography reconstruction? *arXiv preprint arXiv:2506.02093*, 2025. 18

[65] Jie Liu, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for partially labeled organ and pan-cancer segmentation. In *MICCAI 2023 FLARE Challenge*, 2023. 17

[66] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21152–21164, 2023. 5, 6, 7, 17

[67] Jie Liu, Yixiao Zhang, Kang Wang, Mehmet Can Yavuz, Xiaoxi Chen, Yixuan Yuan, Haoliang Li, Yang Yang, Alan Yuille, Yucheng Tang, et al. Universal and extensible language-vision models for organ segmentation and tumor detection from abdominal computed tomography. *Medical Image Analysis*, page 103226, 2024. 5

[68] Xiangde Luo, Wenjun Liao, Jianghong Xiao, Tao Song, Xiaofan Zhang, Kang Li, Guotai Wang, and Shaoting Zhang. Word: Revisiting organs segmentation in the whole abdominal region. *arXiv preprint arXiv:2111.02403*, 2021. 17, 19

[69] Qing Lyu and Ge Wang. Conversion between ct and mri images using diffusion and score-matching models. *arXiv preprint arXiv:2209.12104*, 2022. 18

[70] J. Ma and B. Wang. Fast, low-resource, accurate, and robust organ and pan-cancer segmentation. In *27th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2024)*. Zenodo, 2024. 1, 2, 3

[71] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, et al. Abdomenct-1k: Is abdominal organ segmentation a solved problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 19

[72] Jun Ma, Yao Zhang, Song Gu, Xingle An, Zhihe Wang, Cheng Ge, Congcong Wang, Fan Zhang, Yu Wang, Yinan Xu, et al. Fast and low-gpu-memory abdomen ct organ segmentation: the flare challenge. *Medical Image Analysis*, 82: 102616, 2022. 19

[73] Jiawei Mao, Yuhan Wang, Yucheng Tang, Daguang Xu, Kang Wang, Yang Yang, Zongwei Zhou, and Yuyin Zhou. Medsegfactory: Text-guided generation of medical image-mask pairs. *arXiv preprint arXiv:2504.06897*, 2025. 1

[74] Xiangxi Meng, Yuning Gu, Yongsheng Pan, Nizhuan Wang, Peng Xue, Mengkang Lu, Xuming He, Yiqiang Zhan, and Dinggang Shen. A novel unified conditional score-based generative framework for multi-modal medical image completion. *arXiv preprint arXiv:2207.03430*, 2022. 18

[75] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016. 17

[76] Andriy Myronenko. 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*, pages 311–320. Springer, 2019. 6, 7

[77] Piyush Nathani, Purva Gopal, Nicole Rich, Adam Yopp, Takeshi Yokoo, Binu John, Jorge Marrero, Neehar Parikh, and Amit G Singal. Hepatocellular carcinoma tumour volume doubling time: a systematic review and meta-analysis. *Gut*, 70(2):401–407, 2021. 18

[78] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022. 1, 3

[79] Muzaffer Özbey, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Özturk, Alper Güngör, and Tolga Çukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 2023. 18

[80] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3

[81] Micha Pfeiffer, Isabel Funke, Maria R Robu, Sebastian Bodenstedt, Leon Strenger, Sandy Engelhardt, Tobias Roß, Matthew J Clarkson, Kurinchi Gurusamy, Brian R Davidson, et al. Generating large labeled data sets for laparoscopic image processing tasks using unpaired image-to-image translation. In *Medical Image Computing and Computer Assisted Intervention*, pages 119–127. Springer, 2019. 18

[82] Marion J Pollheimer, Peter Kornprat, Richard A Lindtner, Lars Harbaum, Andrea Schlemmer, Peter Rehak, and Cord Langner. Tumor necrosis is a new promising prognostic factor in colorectal cancer. *Human pathology*, 41(12):1749–1757, 2010. 18

[83] Colin H Richards, Zahra Mohammed, Tahir Qayyum, Paul G Horgan, and Donald C McMillan. The prognostic value of histological tumor necrosis in solid organ malignant disease: a systematic review. *Future oncology*, 7(10): 1223–1235, 2011. 18

[84] Blaine Rister, Darvin Yi, Kaushik Shivakumar, Tomomi Nobashi, and Daniel L Rubin. Ct-org, a new dataset for multiple organ segmentation in computed tomography. *Scientific Data*, 7(1):1–9, 2020. 19

[85] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3, 6

[86] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 17

[87] Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 556–564. Springer, 2015. 19

[88] Younghak Shin, Hemin Ali Qadir, and Ilangko Balasingham. Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance. *IEEE Access*, 6:56007–56017, 2018. 18

[89] Marc C Smaldone, Alexander Kutikov, Brian L Egleston, Daniel J Canter, Rosalia Viterbo, David YT Chen, Michael A Jewett, Richard E Greenberg, and Robert G Uzzo. Small renal masses progressing to metastases under active surveillance: a systematic review and pooled analysis. *Cancer*, 118(4):997–1006, 2012. 18

[90] L Soler, A Hostettler, V Agnus, A Charnoz, J Fasquel, J Moreau, A Osswald, M Bouhadjar, and J Marescaux. 3d image reconstruction for comparison of algorithm database: A patient specific anatomical and medical image database. *IRCAD, Strasbourg, France, Tech. Rep*, 2010. 6, 7

[91] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*, 2021. 18

[92] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. 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*, pages 20730–20740, 2022. 5, 6, 7, 17

[93] Vanya V Valindria, Nick Pawlowski, Martin Rajchl, Ioannis Lavdas, Eric O Aboagye, Andrea G Rockall, Daniel Rueckert, and Ben Glocker. 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)*, pages 547–556. IEEE, 2018. 19

[94] Julia Wolleb, Robin Sandkühler, Florentin Bieder, Philippe Valmaggia, and Philippe C Cattin. Diffusion models for implicit image segmentation ensembles. In *International Conference on Medical Imaging with Deep Learning*, pages 1336–1348. PMLR, 2022. 18

[95] Linshan Wu, Jiaxin Zhuang, Xuefeng Ni, and Hao Chen. Freetumor: Advance tumor segmentation via large-scale tumor synthesis. *arXiv preprint arXiv:2406.01264*, 2024. 18

[96] Yingda Xia, Qihang Yu, Linda Chu, Satomi Kawamoto, Seyoun Park, Fengze Liu, Jieneng Chen, Zhuotun Zhu, Bowen Li, Zongwei Zhou, et al. The felix project: Deep networks to detect pancreatic neoplasms. *medRxiv*, 2022. 1, 6

[97] Tiange Xiang, Yixiao Zhang, Yongyi Lu, Alan Yuille, Chaoyi Zhang, Weidong Cai, and Zongwei Zhou. Exploiting structural consistency of chest anatomy for unsupervised anomaly detection in radiography images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 17

[98] Yutong Xie and Quanzheng Li. Measurement-conditioned denoising diffusion probabilistic model for under-sampled medical image reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 655–664. Springer, 2022. 18

[99] Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. Deeplesion: Automated deep mining, categorization and detection of significant radiology image findings using large-scale clinical lesion annotations. *arXiv preprint arXiv:1710.01766*, 2017. 3

[100] Yijun Yang, Zhao-Yang Wang, Qiuping Liu, Shuwen Sun, Kang Wang, Rama Chellappa, Zongwei Zhou, Alan Yuille, Lei Zhu, Yu-Dong Zhang, et al. Medical world model: Generative simulation of tumor evolution for treatment planning. *arXiv preprint arXiv:2506.02327*, 2025. 17

[101] Jinsung Yoon, Daniel Jarrett, and Mihaela Van der Schaar. Time-series generative adversarial networks. *Advances in neural information processing systems*, 32, 2019. 17

[102] Qihang Yu, Dong Yang, Holger Roth, Yutong Bai, Yixiao Zhang, Alan L Yuille, and Daguang Xu. C2fnas: Coarse-to-fine neural architecture search for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4126–4135, 2020. 5

[103] Zongwei Zhou. *Towards Annotation-Efficient Deep Learning for Computer-Aided Diagnosis*. PhD thesis, Arizona State University, 2021. 1

[104] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11. Springer, 2018. 17

[105] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 39(6): 1856–1867, 2019. 17

[106] Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Medical Image Analysis*, 67:101840, 2021. 5

[107] Zongwei Zhou, Michael B Gotway, and Jianming Liang. Interpreting medical images. In *Intelligent Systems in Medicine and Health*, pages 343–371. Springer, 2022. 1

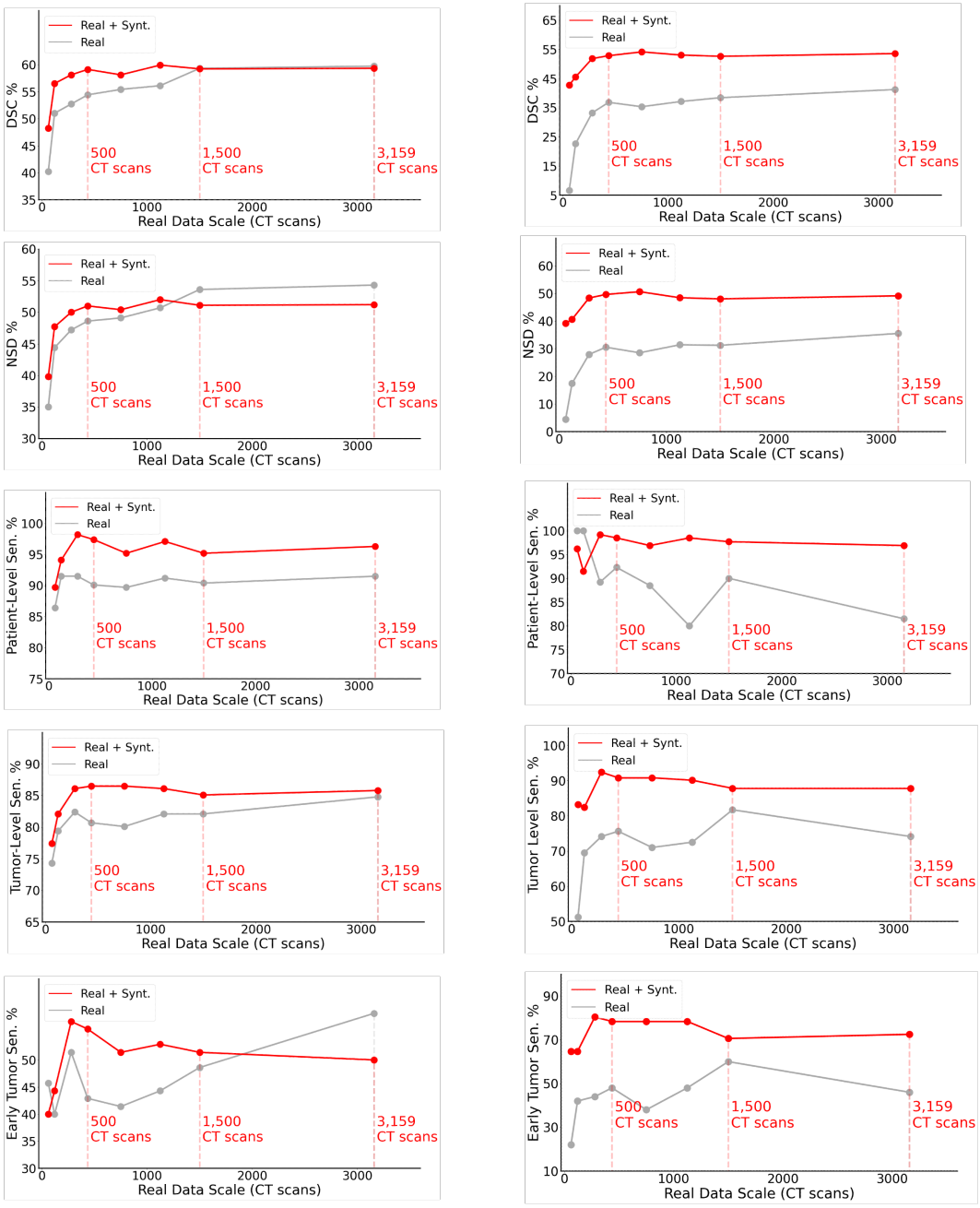
[108] Lingting Zhu, Noel Codella, Dongdong Chen, Zhenchao Jin, Lu Yuan, and Lequan Yu. Generative enhancement for 3d medical images. *arXiv preprint arXiv:2403.12852*, 2024. 18

Scaling Tumor Segmentation: Best Lessons from Real and Synthetic Data

Supplementary Material

This appendix is organized as follows:

A. Best Lesson Proof on proprietary dataset



(a) Best lesson on in-distribution data

(b) Best lesson on out-of-distribution data

Figure 8. **Best lesson proof on proprietary dataset.** Comprehensive experimental results trained on the proprietary dataset show that increasing the scale of real data (gray curve) improves segmentation (DSC and NSD) and detection (patient-level sensitivity, tumor-level sensitivity, and early tumor sensitivity) for both in-distribution and out-of-distribution data. Additionally, augmenting the dataset with an extra 3× synthetic data (red curve) consistently enhances the results. The specific numerical results in this figure can be referenced in the Table 5. Given the substantial GPU requirements, the results were obtained from a single experiment. To reach a more reliable conclusion, we will conduct the experiments at least 10 times.

Scaling with real data.

	#real CT	Patient-level Sen.	Tumor-level Sen.	DSC	NSD	Early Tumor Sen.
Test on in-distribution data	60	86.4	74.3	40.2	35.0	45.7
	120	90.1	79.4	51.0	44.4	40.0
	278	91.5	82.4	52.7	47.2	51.4
	435	90.1	80.7	54.4	48.6	42.9
	750	89.7	80.1	55.4	49.1	41.4
	1125	91.2	82.1	56.1	50.7	44.3
	1500	90.4	82.1	59.3	53.6	48.6
	3159	91.5	84.8	59.7	54.3	58.6
Test on out-of-distribution data	60	100.0	51.2	6.6	4.4	22.0
	120	100.0	69.5	22.6	17.4	42.0
	278	89.2	74.1	33.2	27.9	44.0
	435	92.3	75.6	36.8	30.5	48.0
	750	88.5	71.0	35.3	28.5	38.0
	1125	80.0	72.5	37.1	31.4	48.0
	1500	90.0	81.7	38.4	31.2	60.0
	3159	81.5	74.1	41.2	35.5	46.0

Scaling with real & synthetic data.

	#real CT	Patient-level Sen.	Tumor-level Sen.	DSC	NSD	Early Tumor Sen.
Test on in-distribution data	60	89.7	77.4	48.2	39.8	40.0
	120	94.1	82.1	56.5	47.7	44.3
	278	98.2	86.1	58.1	50.0	57.1
	435	97.4	86.5	59.1	51.0	55.7
	750	95.2	86.5	58.1	50.4	51.4
	1125	97.1	86.1	59.9	52.0	52.9
	1500	95.2	85.1	59.2	51.1	51.4
	3159	96.3	85.8	59.3	51.2	50.0
Test on out-of-distribution data	60	96.2	83.2	42.7	39.2	64.7
	120	91.5	82.4	45.5	40.6	64.7
	278	99.2	92.4	51.8	48.3	80.4
	435	98.5	90.8	52.8	49.6	78.4
	750	96.9	90.8	54.1	50.6	78.4
	1125	98.5	90.1	53.0	48.4	78.4
	1500	97.7	87.8	52.6	48.0	70.6
	3159	96.9	87.8	53.5	49.1	72.5

Table 5. **Best Lesson Proof on proprietary dataset.** The proprietary dataset comprises a total of 5,176 CT scans, which include scans of patients with pancreatic tumors as well as healthy scans without pancreatic tumors. We utilized 3,159 scans for training, while the remaining 2,017 were allocated for testing within the same distribution. For the out-of-distribution dataset, we selected the Panorama dataset. Detailed information regarding the dataset split can be found in § C. For the segmentation model, we employed the SegResNet model based on the MONAI codebase for training and assessed the tumor segmentation and detection results using the DSC, NSD, and sensitivity metrics.

B. Related works

B.1. AI Development on Real Tumors

AI algorithms. Tumor detection and segmentation have been long-standing problems in medical image analysis. To achieve deliverable results, many recent works leverage state-of-the-art deep learning technology [43, 92].

The U-Net architecture [86] has been widely adopted in medical image analysis. Over the years, numerous well-designed networks have been proposed to improve the U-Net architecture, including UNet++ [104, 105], TransU-Net [9], UNETR [32], Swin-UNETR [31], and many others [8, 10, 16, 75]. While these methods have demonstrated remarkable performance in tumor detection and segmentation, they typically rely on a significant number of annotations. The process of annotating real tumors is not only time-consuming but also requires extensive medical expertise. Sometimes, it needs the assistance of radiology reports [4, 5] or is even impossible to obtain the annotation [6, 42, 97, 100]. Therefore, the use of synthetic tumors emerges as a promising solution.

Liu et al. [66] integrate text embeddings derived from Contrastive Language-Image Pre-training (CLIP) into segmentation models, effectively capturing anatomical relationships and enabling the model to learn structured feature embeddings across multiple organ and tumor types. With pre-training on large-scale CT scans with per-voxel annotations for 25 anatomical structures and seven tumor types, Li et al [60] has developed a suite of models demonstrating robust transfer learning capabilities across various downstream organ and tumor segmentation tasks.

Preexisting public datasets have made significant contributions to the advancement of AI in tumor detection [61]. We summarizes key characteristics of existing public datasets for organ and tumor segmentation in table 1, categorized into those with and without tumor labels. Datasets such as LiTS [6] and KiTS [36] provide essential tumor labels but are limited with regard to size and variety, with 131 and 489 scans, respectively, and fewer hospitals contributing data (7 for LiTS and 1 for KiTS). Larger datasets like FLARE23 [65] include 2,200 scans and span contributions from 30 hospitals, yet they focus on a single organ and provide no explicit tumor-specific labels. Similarly, datasets without tumor labels, such as WORD [68] and AMOS22 [46], are useful for broader anatomical segmentation tasks but lack tumor-specific annotations. In contrast, AbdomenAtlas 2.0 distinguishes itself by offering the most extensive dataset to date, with 10,136 scans, 4,700K slices, and 13,223 tumors annotated across multiple organs, including rarer tumor types like esophagus and uterus. The dataset incorporates data from 89 hospitals across a wide range of countries, providing unprecedented diversity and comprehensiveness for multi-organ tumor research.

B.2. AI Development on Synthetic Tumors

Tumor synthesis enables the generation of artificial tumors in medical images, aiding in the training of AI models for tumor detection and segmentation [14, 48, 101]. Synthetic tumors become particularly valuable when acquiring per-voxel annotations of real tumors is challenging, such as in the early stages of tumor development. There are several advantages of synthetic tumors over real tumors.

Quality Control: Synthetic data allows for the control of specific variables and the introduction of desired diversity into the dataset. Real-world datasets often suffer from imbalances, such as an overrepresentation of certain demographics or tumor stages. Synthetic data can be generated to balance these datasets, ensuring that machine learning models are trained on a comprehensive and representative sample of data. For rare cancers, collecting enough patient data is particularly difficult. Synthetic data can help augment these limited datasets, enabling the development of more robust and accurate models for rare cancer types. Additionally, synthetic data can be used to simulate hard cases that are difficult to capture in real-world data. Researchers can rapidly iterate and refine their models, leading to faster advancements in tumor detection, diagnosis, and treatment.

Privacy and Ethical Considerations: One of the major advantages of synthetic data is that it can be used without compromising patient privacy. Since synthetic data is not directly tied to any real individual, it eliminates the risk of exposing sensitive patient information. By using synthetic data, researchers can bypass ethical dilemmas associated with real patient data, such as the need for patient consent and the risk of data breaches.

Synthetic tumors can be used in aiding AI models for tumor detection and segmentation, particularly in situations where detailed annotations are scarce [14, 19]. Therefore, an effective and universally applicable tumor synthesis approach is urgently needed to accelerate the development of tumor detection and segmentation methods.

Tumor development is intricately regulated by biological mechanisms at various scales. Tumors, which arise from DNA mutations in a single cell and represent genetic disorders [55], undergo complex growth processes. Mutated cells lead to uncontrolled proliferation, which can be benign or malignant [24]. Differences between benign and malignant tumors include

growth rate and invasiveness [55]. Malignant tumors tend to exhibit larger final sizes and faster growth rates compared to benign lesions [50]. Additionally, slow tumor growth rates have been associated with low malignant potential [15, 89]. These patterns have also been observed in several studies [26, 77]. Malignant tumors usually invade surrounding tissues, while benign tumors typically remain confined to their original sites. Moreover, even slowly growing malignant tumors can invade surrounding tissues [52], leading to blurry boundaries between tumors and adjacent tissues. Therefore, it is necessary to design Accumulation and Growth rules to simulate these features. Tumor necrosis, a form of cell death, indicates a worse prognosis [82, 83]. Histologically, necrosis is caused by hypoxia resulting from rapid cell proliferation surpassing vascular supply [39], presenting as non-enhancing irregular areas in CT images [22]. Hu et al. [41] developed a program that integrates medical knowledge to generate realistic liver tumors. However, these models are generally organ-specific and require adaptation to work with other organs. Lai et al. [56] proposed a framework that leverages cellular automata to simulate tumor growth, invasion, and necrosis, enabling realistic synthetic tumor generation across multiple organs.

Generative models have been effectively utilized in the medical field for tasks like image-to-image translation [69, 74, 79, 81], reconstruction [64, 91, 98], segmentation [11, 21, 53, 94], and image denoising [25]. Utilizing advanced generative models to synthesize various tumors is also a promising direction [27, 30, 95, 108]. Shin et al. [88] advanced detection by generating synthetic abnormal colon polyps using Conditional Adversarial Networks. Chen et al. [12] employed a diffusion model that capitalizes on similarities in early-stage tumor imaging for cross-organ tumor synthesis. Wu et al. [95] employs an adversarial-based discriminator to automatically filter out the low-quality synthetic tumors to improve tumor synthesis. Guo et al. [27] incorporates ControlNet to process organ segmentation as additional conditions to guide the generation of CT images with flexible volume dimensions and voxel spacing.

C. Implementation Details

C.1. Dataset Composition

AbdomenAtlas 2.0 components	# of scans	annotated tumor (original)	annotators
Public CT in AbdomenAtlas 2.0 (AbdomenAtlas1.1)	9,901	liver, pancreas, kidney, colon	human & AI
CHAOS [2018] [link]	20	-	human
BTCV [2015] [link]	47	-	human
Pancreas-CT [2015] [link]	42	-	human
CT-ORG [2020] [link]	140	-	human & AI
WORD [2021] [link]	120	-	human
LiTS [2019] [link]	130	liver	human
AMOS22 [2022] [link]	200	-	human & AI
KiTS [2023] [link]	489	kidney	human
AbdomenCT-1K [2021] [link]	1,000	-	human & AI
MSD-CT [2021] [link]	945	liver, pancreas, colon	human & AI
FLARE'23 [2022] [link]	4,100	-	human & AI
Abdominal Trauma Det [2023] [link]	4,711	-	-
Private CT in AbdomenAtlas 2.0	233	liver, pancreas, kidney, colon, esophagus, uterus	human & AI

Table 6. **Dataset composition of AbdomenAtlas 2.0.** Our AbdomenAtlas 2.0 comprises two components: CT scans from the public AbdomenAtlas 1.1 dataset and CT scans from a private source, totaling 10,134 tumor-annotated CT scans, with additional scans expected from various sources. Note that, for CT scans from AbdomenAtlas 1.1 dataset, we fully annotate six tumor types for each CT scan.

C.2. Comparative Models

The code for the Comparative Model is implemented in Python using MONAI and nnU-Net framework.

nnU-Net Framework. nnU-Net serves as a framework for the automatic configuration of AI-driven semantic segmentation pipelines. When presented with a new segmentation dataset, it extracts pertinent metadata from the training cases to automatically determine its hyperparameters. It has withstood the test of time and continues to deliver state-of-the-art results. nnU-Net effectively illustrates that meticulously configuring and validating segmentation pipelines across a diverse range of segmentation tasks can yield a remarkably powerful algorithm.

We implement UNETR, Swin UNETR, nnU-Net, ResEncM, and STU-Net using the nnU-Net framework. The orientation of CT scans is adjusted to specific axcodes. Isotropic spacing is employed to resample each scan, achieving a uniform voxel size of $1.5 \times 1.5 \times 1.5mm^3$. Additionally, the intensity in each scan is truncated to the range [-175, 250] and then linearly normalized to [0, 1]. During training, we crop random fixed-sized $96 \times 96 \times 96$ regions, selecting centers from either a foreground or background voxel according to a pre-defined ratio. Furthermore, the data augmentation during training adheres to the default strategies outlined in the nnU-Net framework. All models are trained for 1000 epochs, with each epoch consisting of 250 iterations. Besides, we utilize the SGD optimizer with a base learning rate of 0.01, and the batch size is defined as 2. During inference, we utilize the test time augmentation by following the default implementations in nnU-Net framework. Besides, we use the sliding window strategy by setting the overlapping area ratio to 0.5.

MONAI Framework. MONAI (Medical Open Network for AI) is an open-source framework that supports AI in healthcare. Built on PyTorch, it offers a comprehensive set of tools for configuring, training, inferring, and deploying medical AI models. We implement SegResNet, Universal Model, and Suprem utilizing the MONAI framework. Since different methods have varying hyperparameter settings, we trained and tested the models exactly according to the original hyperparameters specified in the corresponding papers.

D. Visual Real Examples in AbdomenAtlas 2.0

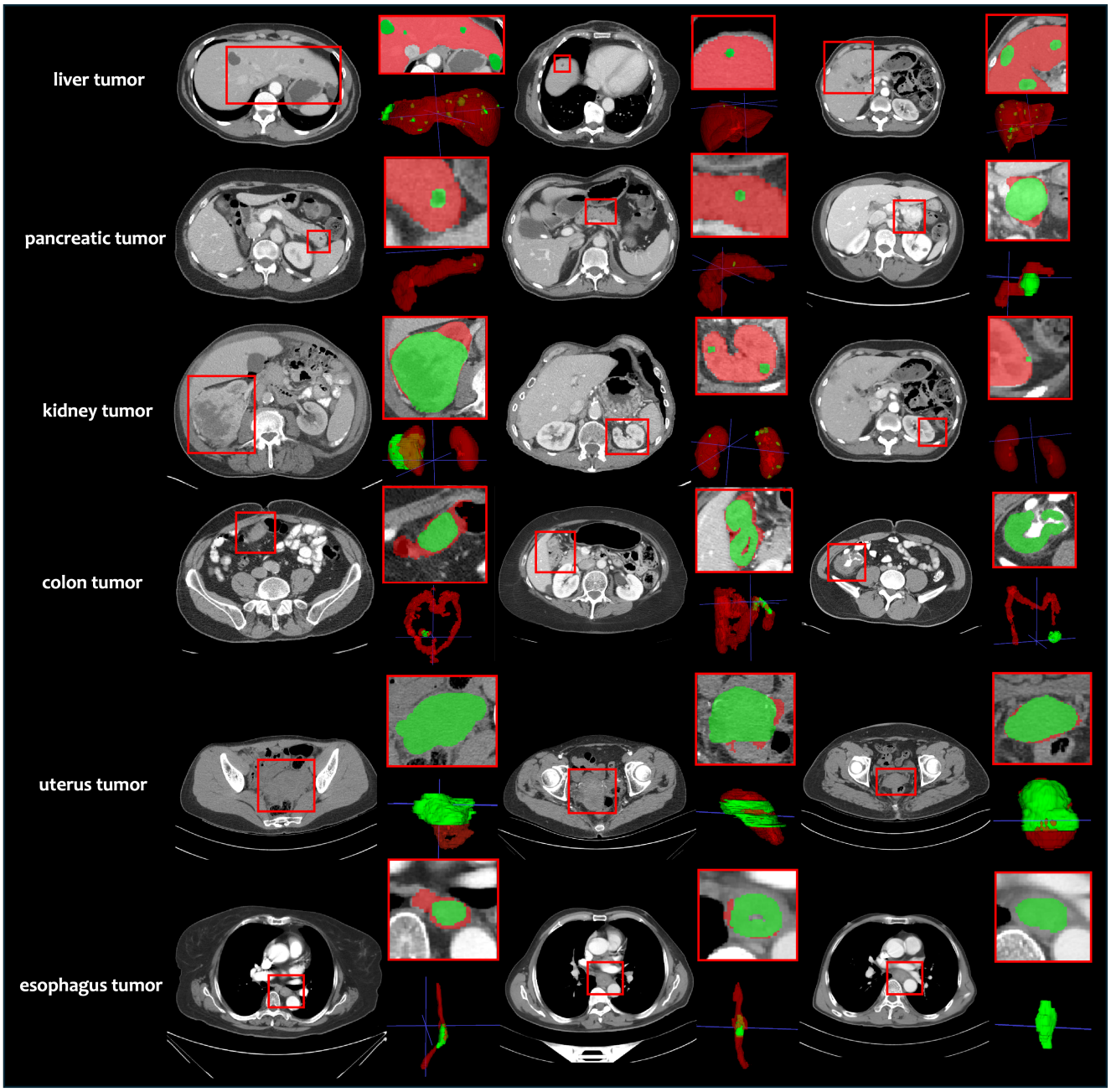


Figure 9. **Visual examples of six tumor types annotated in AbdomenAtlas 2.0.** AbdomenAtlas 2.0 features a diverse distribution across various tumor stages and sizes. These comprehensive, high-quality tumors, accompanied by per-voxel annotations, significantly improve the performance of AI models, both on in-distribution and out-of- distribution data. (Figure 10).

E. More Results: Best Lesson from Real Data

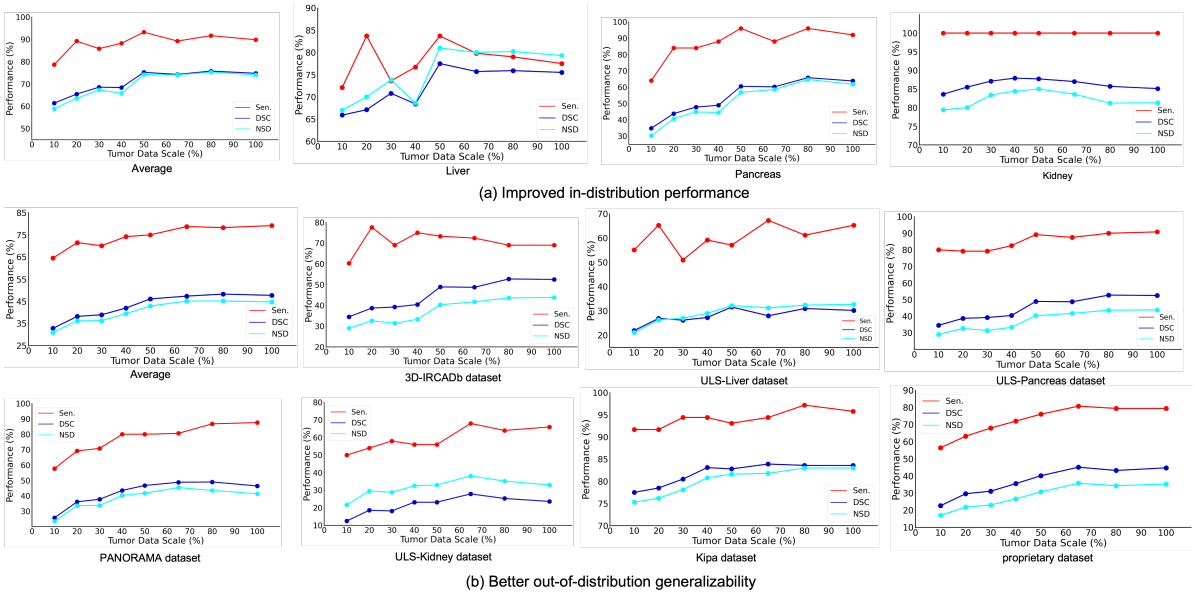


Figure 10. **Best Lesson from Real Data: Results on in-distribution and out-of-distribution data.** (a): while increasing data scale initially enhances in-distribution performance across various metrics (sensitivity, DSC, and NSD), it eventually plateaus. Notably, certain organ types, such as the Liver and Kidney, exhibit a decline in performance at the largest scales. (b): In contrast, the scaling trends observed in out-of-distribution datasets demonstrate consistent improvements in specific datasets (e.g., 3D-IRCADb, ULS-Pancreas) without reaching a plateau, indicating that larger data volumes may enhance generalizability. These results relate to the data-scaling lesson in §1 (1,500 if with real data only). Larger datasets are needed for effective out-of-distribution generalizability.

F. More Results: Best Lesson for generalizability

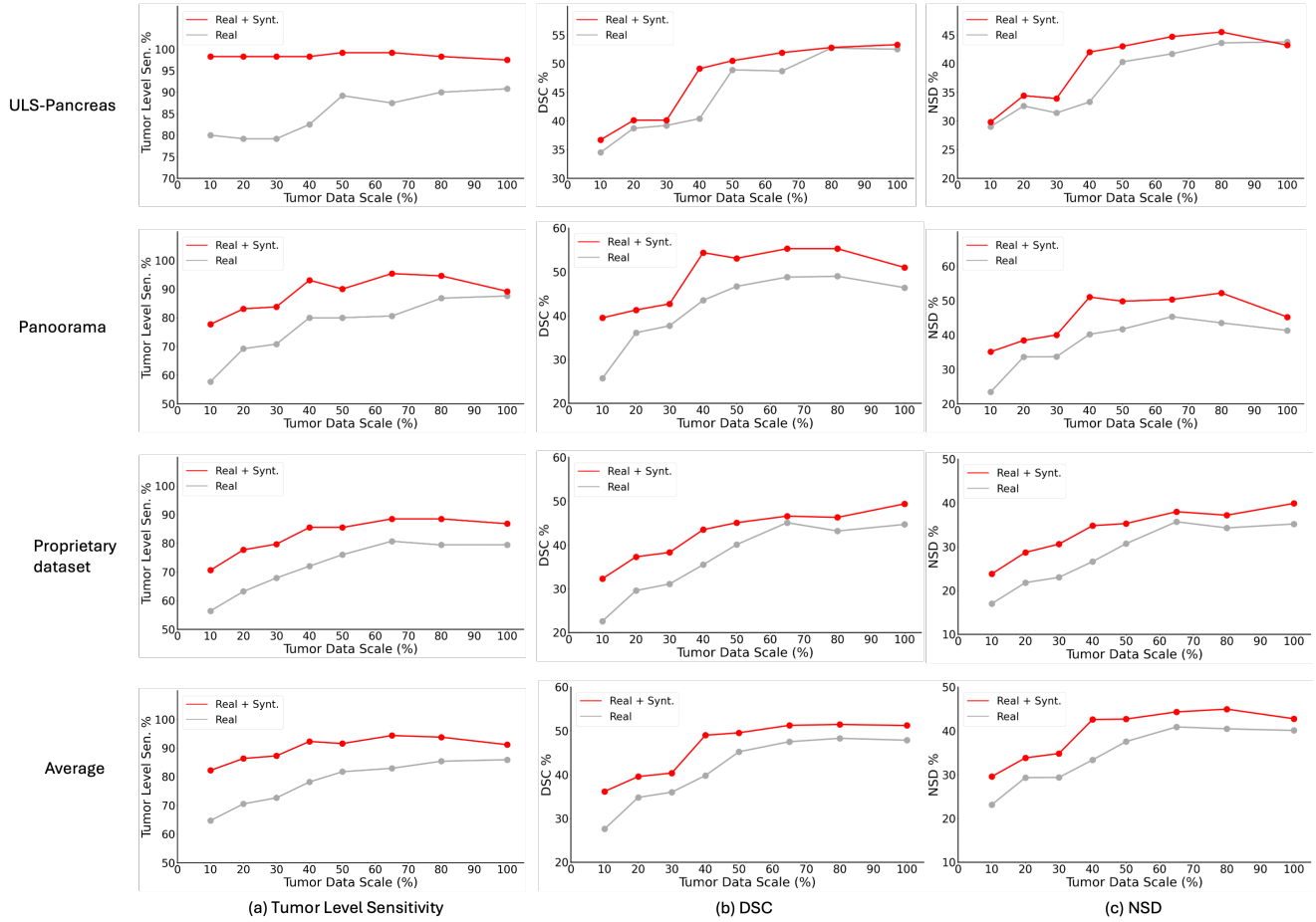


Figure 11. **Best lesson for pancreatic tumors.** Integrating real and synthetic data, compared to using real data alone, consistently improves generalizable performance in sensitivity, DSC, and NSD across various scenarios and data scales. These results underscore the benefits of this combination in enhancing the accuracy of pancreatic tumor analysis.

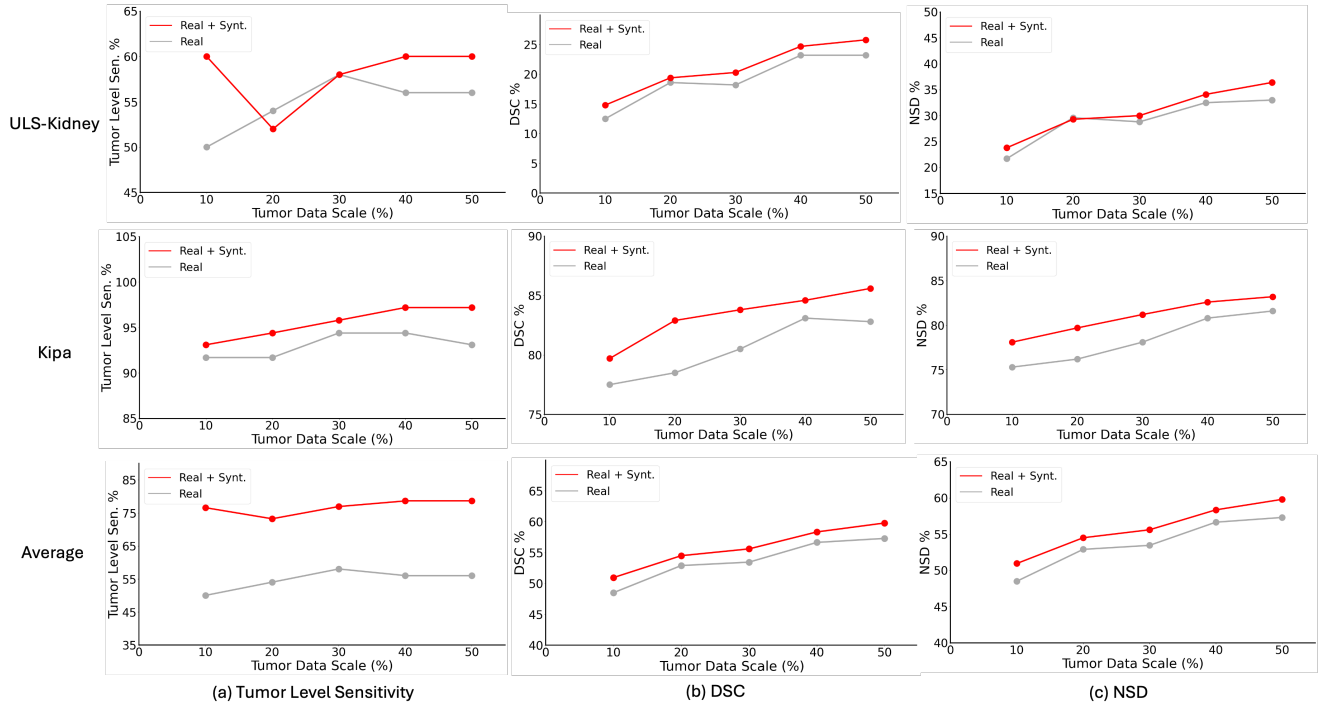


Figure 12. **Best lesson for kidney tumors.** Combining real and synthetic data consistently enhances generalizable performance in sensitivity, DSC, and NSD across various scenarios and data scales, highlighting its effectiveness in improving kidney tumor diagnosis accuracy.