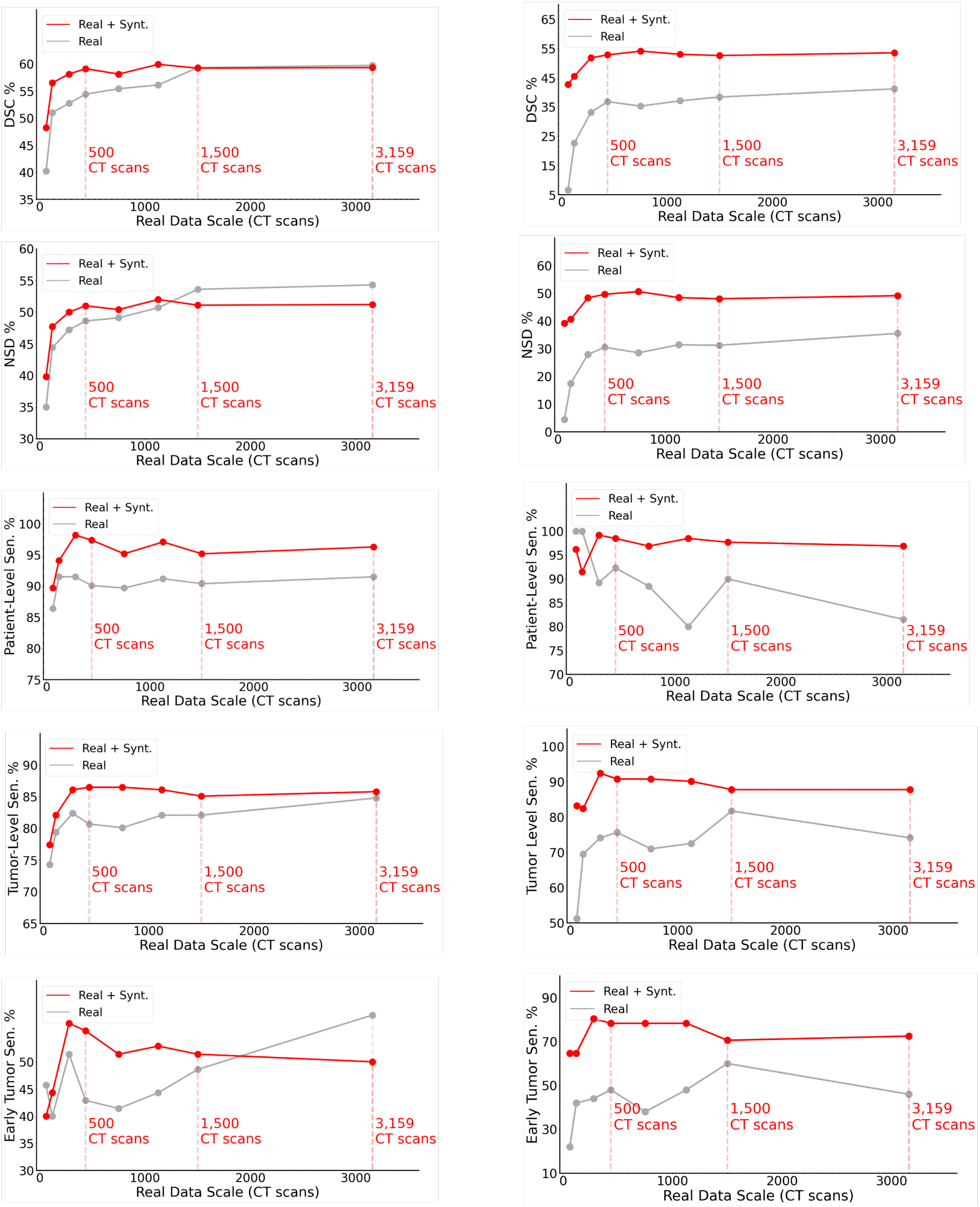
Scaling Tumor Segmentation:
Best Lessons from Real and Synthetic Data

Supplementary Material

This appendix is organized as follows:

- § A provides comprehensive results with scaled real data with proprietary dataset and synthetic data.
- § B provides comprehensive related works.
  - B.1: AI Development on Real Tumors
  - B.2: AI Development on Synthetic Tumors
- § C provides implementation details for Tumor Genesis and comparative models.
  - C.1: details of public and private datasets used in AbdomenAtlas 2.0.
  - C.2: implementation details of comparative models
- § D provides more visual examples from AbdomenAtlas 2.0.
- § E presents additional results on the key insights gained from scaling real tumor data.
- § F presents additional results on the key insights gained from scaling real and synthetic tumor data.

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

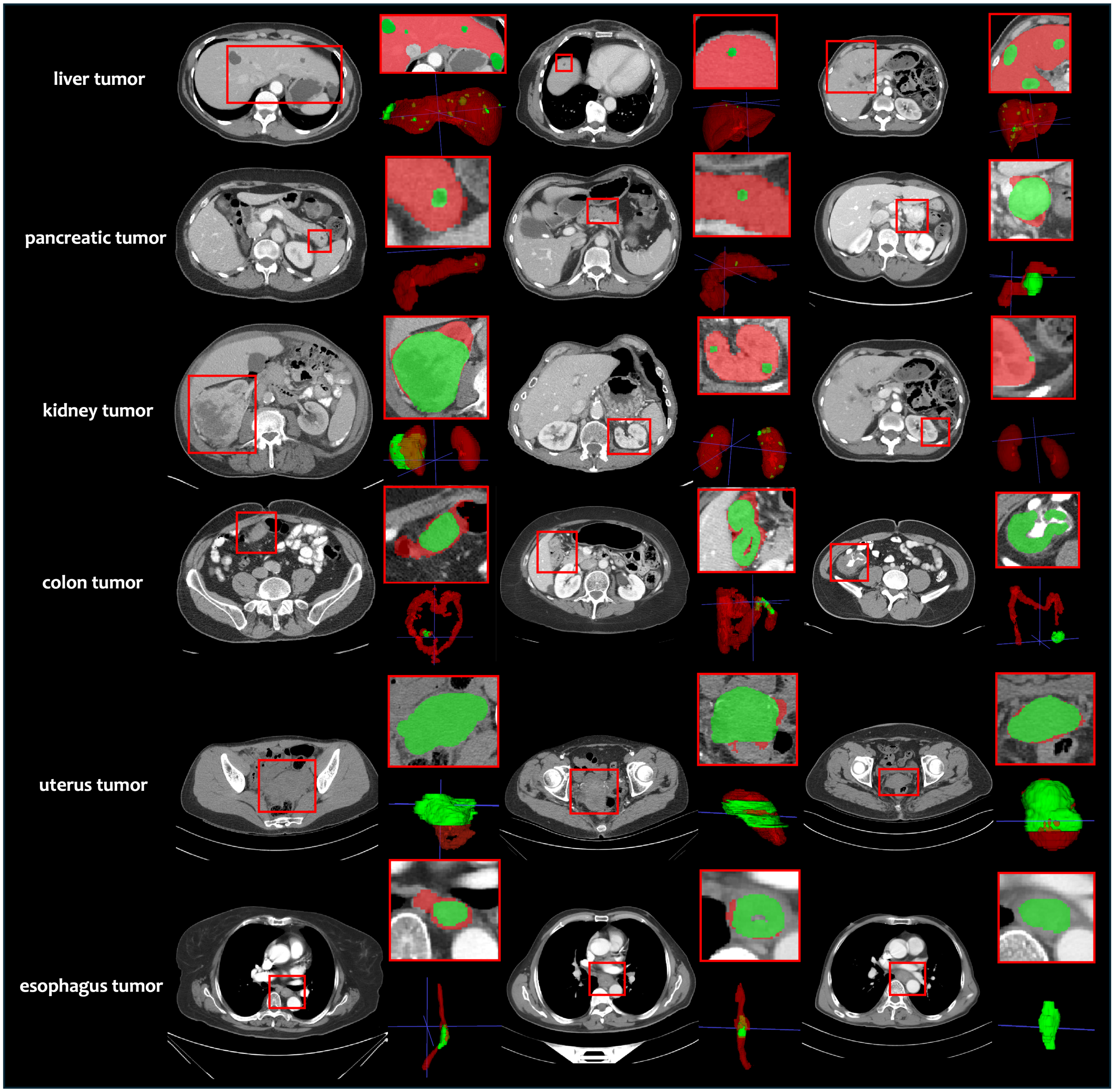
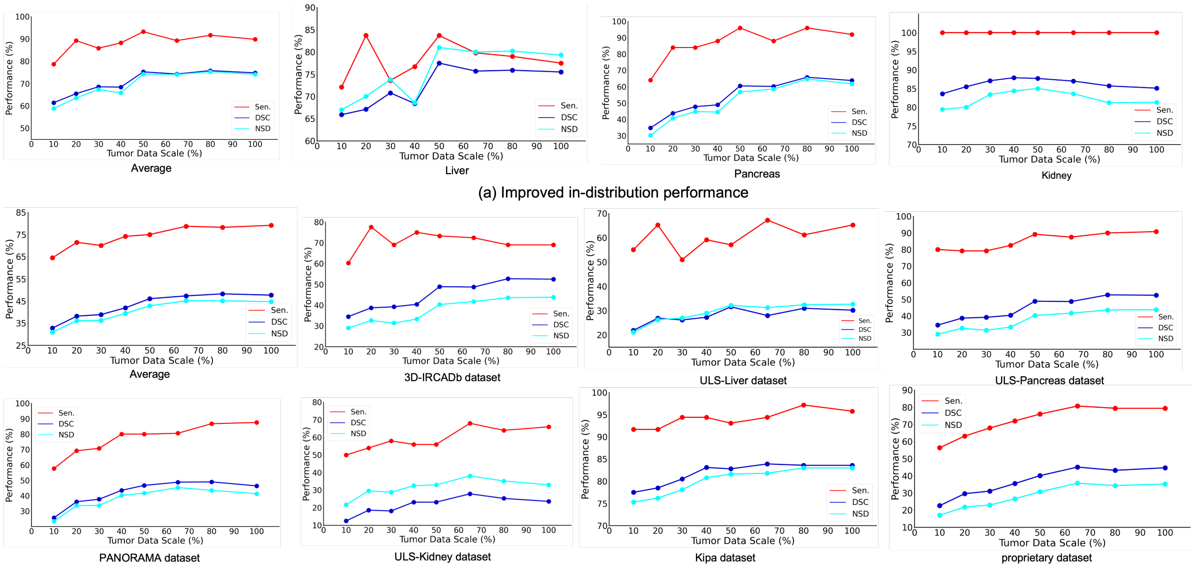


Figure 9. **Visual examples of six tumor types annotated in AbdomenAtlas 2.0.** AbdomenAtlas 2.0 features a diverse distribution across various tumor stages and sizes. These comprehensive, high-quality tumors, accompanied by per-voxel annotations, significantly improve the performance of AI models, both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

E. More Results: Best Lesson from Real Data



Average

Liver

Pancreas

Kidney

(a) Improved in-distribution performance

Average

3D-IRCADb dataset

ULS-Liver dataset

ULS-Pancreas dataset

PANORAMA dataset

ULS-Kidney dataset

Kipa dataset

proprietary dataset

(b) Better out-of-distribution generalizability

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

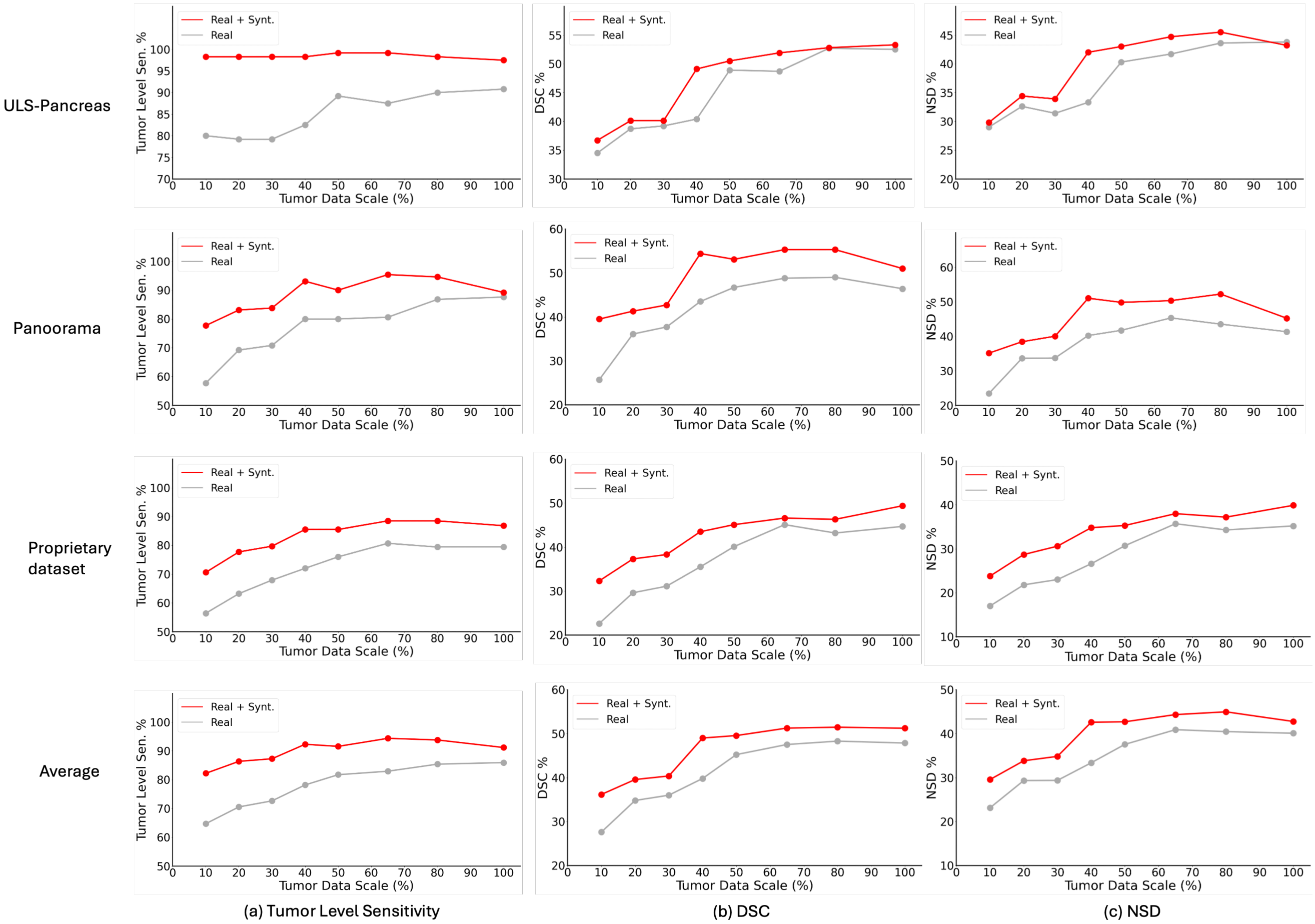
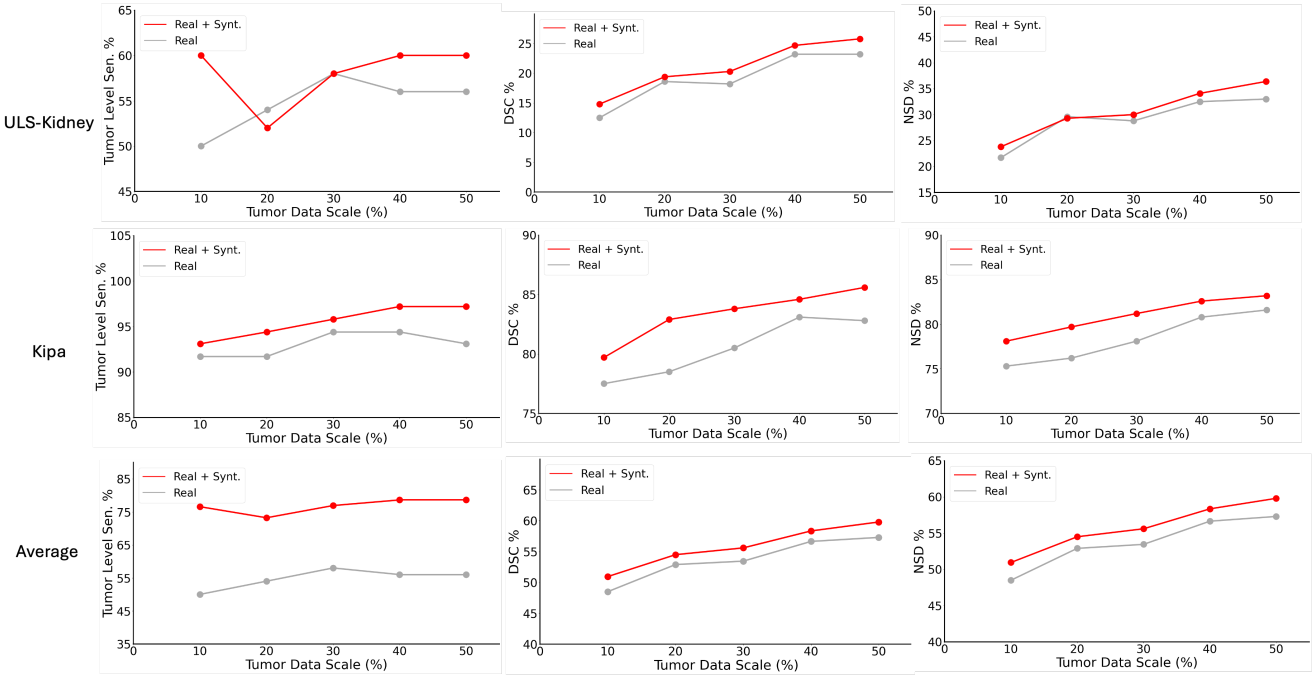


Figure 11. **Best lesson for pancreatic tumors.** Integrating real and synthetic data, compared to using real data alone, consistently improves generalizable performance in sensitivity, DSC, and NSD across various scenarios and data scales. These results underscore the benefits of this combination in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



(a) Tumor Level Sensitivity

(b) DSC

(c) NSD

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.