# PanTS: The Pancreatic Tumor Segmentation Dataset

**Wenxuan Li**[1]* **Xinze Zhou**[1]* **Qi Chen**[1]* **Tianyu Lin**[1] **Pedro R. A. S. Bassi**[1,2,3] **Szymon Płotka**[4] **Jarosław B. Ćwikła**[5,6] **Xiaoxi Chen**[7] **Chen Ye**[8] **Zheren Zhu**[9,10] **Kai Ding**[11] **Heng Li**[11] **Kang Wang**[9] **Yang Yang**[9] **Yucheng Tang**[12] **Daguang Xu**[12] **Alan L. Yuille**[1] **Zongwei Zhou**[1]†

[1]Department of Computer Science, Johns Hopkins University
[2]Department of Pharmacy and Biotechnology, University of Bologna
[3]Center for Biomolecular Nanotechnologies, Istituto Italiano di Tecnologia
[4]Faculty of Mathematics and Computer Science, Jagiellonian University
[5]Department of Cardiology and Internal Medicine, University of Warmia and Mazury
[6]Diagnostic and Treatment Center Gammed
[7]Department of Bioengineering, University of Illinois Urbana-Champaign
[8]Department of General Surgery, Peking University Third Hospital
[9]Department of Radiology & Biomedical Imaging, University of California, San Francisco
[10]Department of Bioengineering, University of California, Berkeley
[11]Department of Radiation Oncology, Johns Hopkins School of Medicine
[12]NVIDIA

Code, Models & Data: https://github.com/MrGiovanni/PanTS

## Abstract

PanTS is a large-scale, multi-institutional dataset curated to advance research in pancreatic CT analysis. It contains 36,390 CT scans from 145 medical centers, with expert-validated, voxel-wise annotations of over 993,000 anatomical structures, covering pancreatic tumors, pancreas head, body, and tail, and 24 surrounding anatomical structures such as vascular/skeletal structures and abdominal/thoracic organs. Each scan includes metadata such as patient age, sex, diagnosis, contrast phase, in-plane spacing, slice thickness, etc. AI models trained on PanTS achieve significantly better performance in pancreatic tumor detection, localization, and segmentation than those trained on existing public datasets. Our analysis indicates that these gains are directly attributable to the 16× larger-scale tumor annotations and indirectly supported by the 24 additional surrounding anatomical structures. As the largest and most comprehensive resource of its kind, PanTS offers a new benchmark for developing and evaluating AI models in pancreatic CT analysis.

## 1 Introduction

Pancreatic cancer is the third leading cause of cancer-related death in the U.S. in both men and women combined [56, 55, 65]. Yet despite its clinical importance, early detection remains a major challenge due to the absence of disease-specific symptoms and the incidental nature of abdominal imaging [48]. Consequently, 80–85% of pancreatic tumors are diagnosed at advanced stages, when treatment options are limited and prognosis is poor [68]. In contrast, early-stage tumors are associated with markedly better outcomes, emphasizing the urgent need for earlier identification [71].

*Equal contribution.
†Correspondence to: Zongwei Zhou (ZZHOU82@JH.EDU)

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

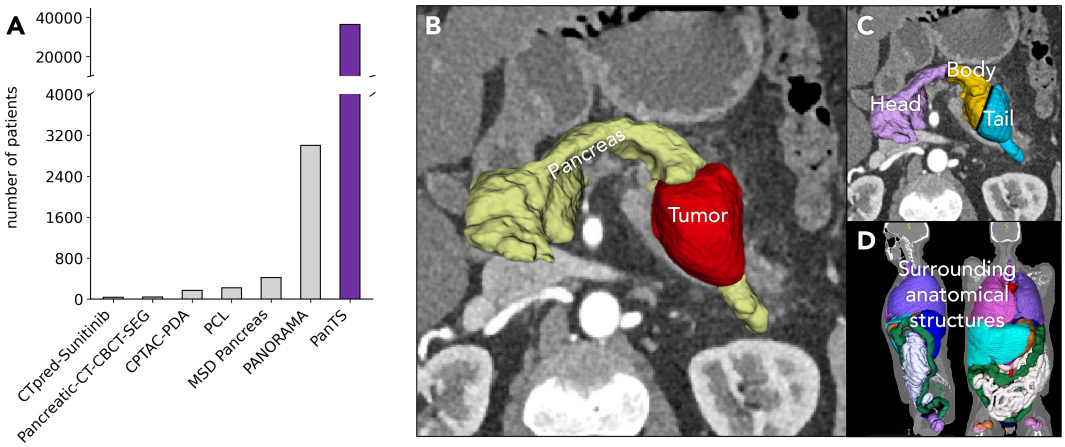


Figure 1: **Dataset characteristics and visualization. A.** PanTS comprises 36,390 CT scans collected from 145 medical centers, paired with expert-validated voxel-wise annotations, 16× larger than the biggest public dataset (i.e., PANORAMA [4]) to date. **B–C.** The dataset includes detailed annotations for pancreatic tumors, pancreas, and its head, body, and tail, enabling spatially aware tumor localization. **D.** Twenty-four surrounding anatomical structures are voxel-wise annotated to provide rich spatial context, including key vessels, ducts, and organs critical for tumor detection, resectability assessment, and radiotherapy planning.

Computed tomography (CT), especially with contrast enhancement, is the primary modality for evaluating pancreatic abnormalities [16]. Retrospective studies have shown that early radiographic signs—such as ductal dilation or focal atrophy—can appear months before clinical diagnosis, but often go undetected [24, 17, 33]. However, these indicators are frequently missed in clinical practice, particularly when scans are acquired for unrelated reasons [57, 61]. Pancreatic tumors in CT scans are highly heterogeneous in shape, size, location, and radiologic appearance [50].

Recent advances in AI have shown promise in automating the detection and localization of pancreatic tumors in CT scans [11, 35, 39]. However, most publicly available models are trained on small, homogeneous datasets and fail to generalize to diverse clinical settings. This shortcoming reflects a fundamental data limitation: the pancreas is a small, anatomically intricate organ embedded among critical vessels, ducts, and adjacent structures, making comprehensive annotation and assessment particularly challenging [25, 38, 36]. Accurate analysis of pancreatic tumors depends not only on identifying the tumor itself but also on understanding its anatomical context.

To address this limitation, we present the Pancreatic Tumor Segmentation Dataset (PanTS)—the largest and most comprehensive dataset to date for pancreatic CT analysis[3]. PanTS comprises 36,390 CT scans from 145 medical centers. Each scan is paired with metadata, including patient age, sex, contrast phase, diagnosis, in-plane spacing, and slice thickness. Importantly, PanTS includes over 993,000 expert-validated voxel-wise annotations (examples in Figure 1), covering:

- Pancreatic tumors along with pancreas head, body, and tail, to enable tumor detection, localization, and segmentation. We find that increasing the number of annotated tumors *directly* improves AI performance on out-of-distribution datasets (Figure 5). To this end, a team of 23 radiologists have produced voxel-wise tumor annotations in each CT scan to support effective AI training at scale.
- Twenty-four surrounding anatomical structures (e.g., superior mesenteric artery, bile ducts; full list in §3) are annotated to enable comprehensive tumor analysis. Joint training on tumors and nearby structures *indirectly* enhances AI performance by reducing false positives and providing rich anatomical context (Figure 6). Feature analysis reveals that models trained with both tumor and anatomical structure labels learn more discriminative and separable representations, allowing for more precise tumor detection and segmentation.

[3]PanTS is not intended for direct clinical decision-making or real-time diagnosis.

With its large scale, diversity, and anatomical detail, PanTS sets a new benchmark for AI development in pancreatic CT analysis. It includes 9,901 publicly available training scans (non-commercial license) and 26,489 test scans reserved for third-party evaluation. This setup follows best practices in medical AI benchmarking [47, 7, 6, 37], ensuring fair and reproducible comparisons. We also release a strong baseline model, nnU-Net, alongside the dataset. This baseline model ranked Top-1 in the official Medical Segmentation Decathlon (MSD) Leaderboard.

## 2 Related Datasets & Our Contribution

### 2.1 Pancreas and Other Organ Datasets

Several public datasets have advanced multi-organ segmentation in abdominal CT, including BTCV [34] (50 CTs, 13 classes, 1 center), CHAOS [31] (40 CTs, 4 class, 1 center), AMOS22 [28] (500 CTs, 15 classes, 2 centers), WORD [43] (150 CTs, 16 classes, 1 center), and AbdomenCT-1K [44] (1,112 CTs, 4 classes, 12 centers). These datasets typically target general abdominal structures or liver segmentation, with limited diversity in institution count ($\leq$12 centers) and relatively modest dataset sizes. TotalSegmentator [64] is one of the most ambitious efforts to date, offering 1,228 CT scans across 117 classes from a single source. However, its focus remains on broad anatomic structure segmentation and lacks dedicated design for oncologic applications.

***Limitation:*** While these datasets are useful for general anatomical segmentation, they are not specifically designed for pancreatic tumor analysis. None of them provides voxel-wise annotations of important pancreatic substructures, such as the head, body, and tail of the pancreas, the superior mesenteric artery, pancreatic duct, common bile duct, celiac artery, and duodenum. These annotations are essential for surgical decision-making, tumor staging, and accurate assessment of tumor invasion and resectability. Reference organs such as the liver, spleen, kidneys, adrenal glands, aorta, and postcava are either inconsistently labeled or absent [40, 41, 29, 72, 70, 59]. Furthermore, distal anatomical landmarks, including the lungs, femurs, bladder, and prostate, which are important for spatial orientation and radiotherapy planning, are rarely included.

***Our Contribution:*** PanTS addresses these limitations by offering voxel-wise annotations for 27 clinically meaningful structures selected specifically to support pancreatic tumor analysis. These include voxel-wise annotations of the pancreas head, body, and tail, and 24 surrounding anatomical structures crucial for spatial reasoning, proximity assessment, and downstream clinical workflows such as radiotherapy planning and vessel invasion analysis. With 36,390 CT scans from 145 global medical centers, PanTS is not only the largest organ segmentation dataset available, but also the most diverse—offering over 3$\times$ more institutional representation and over 7$\times$ more data than leading datasets like AbdomenCT-1K [44] or AMOS22 [28].

### 2.2 Pancreatic and Other Tumor Datasets

Tumor segmentation datasets have historically focused on more common cancers and organs. For instance, liver tumors are supported by datasets like LiTS [10] (201 CTs, 7 centers), HCC-TACE-Seg [49] (105 CTs), and MSD Liver [6] (201 CTs); colorectal tumors by Stagell-Colorectal-CT [60] (230 CTs); kidney tumors by TCGA-KIRC [3] (267 CTs) and KiTS23 [21] (599 CTs); and lung tumors by MSD Lung [6] (96 CTs). Large-scale efforts such as FLARE'23 [46] (4,500 CTs, 14 classes, more than 50 centers) and autoPET [2] (1,214 CTs, 1 class) target pan-cancer analysis but lack pancreas-specific detail or annotations of relevant anatomical structures.

***Limitation:*** Pancreatic tumor datasets, in comparison, remain scarce and small in scale [8, 14, 9, 15]. NIH Pancreas-CT [1] (82 CTs), Pancreatic-CT-CBCT-SEG [23] (40 CTs), and CPred-Sunitinib-panNET [13] (38 CTs) are all limited to single centers and focus on narrow tumor types or clinical scenarios. PANORAMA [4] (2,238 CTs, 6 classes, 7 centers) is a major step forward, offering voxel-wise annotations for pancreatic ductal adenocarcinoma (PDAC) and associated structures such as ducts and vessels. However, it does not provide annotations for other types of pancreatic tumors, which causes issue in evaluation as discussed in §4.

***Our Contribution:*** PanTS is the largest and most comprehensive publicly available dataset for pancreatic tumor segmentation, offering over 16$\times$ more annotated CT scans than PANORAMA and spanning over 20$\times$ more medical centers. In addition to voxel-wise annotations of pancreatic tumors, PanTS provides segmentation of the pancreas head, body, and tail, enabling precise tumor

localization and region-aware staging. The dataset supports a full pipeline of clinically relevant tasks—tumor detection, segmentation, staging, resectability assessment, and surgical planning—by also including 24 surrounding anatomical structures critical for evaluating tumor involvement of vessels and adjacent organs. No existing dataset provides this combination of scale, diversity, and task-aligned anatomical detail.

## 3 PanTS: The Pancreatic Tumor Segmentation Dataset

PanTS comprises 36,390 CT scans with precise per-voxel annotations of pancreatic tumors, pancreas head, body, and tail, along with 24 surrounding structures (i.e., pancreas, superior mesenteric artery, pancreatic duct, celiac artery, common bile duct, veins, aorta, gall bladder, left and right kidneys, liver, postcava, spleen, stomach, left and right adrenal glands, bladder, colon, duodenum, left and right femurs, left and right lungs, and prostate). Sourced from 145 centers, this dataset includes imaging metadata such as patient sex, age, contrast phase, diagnosis, spacing, and scanner details.

We split the PanTS into a training set of 9,901 cases (27%) and a test set of 26,489 cases (73%), both consisting of abdominal CT scans. For public reproducibility, the training set is further split into 9,000 cases for model development and 901 cases as an official public test set [4]. Detailed dataset characteristics are summarized in Table 1. The data and annotation are licensed as CC BY-NC-SA. We have released the training set to The PanTS Huggingface Website, and the test set is preserved for third-party evaluation.

### 3.1 Dataset Diversity

The PanTS dataset comprises a broad spectrum of pancreatic tumor types, including pancreatic ductal adenocarcinoma, pancreatic neuroendocrine tumors (PNETs), pancreatic cystic neoplasms, and cystic non-neoplastic lesions. These entities exhibit heterogeneous imaging characteristics in terms of size, morphology, attenuation, and texture. The CT scans are abdominal images obtained using varying contrast phases, scanner models, and imaging protocols. The dataset also contains real-world imaging artifacts, such as metal-induced streaks, contributing to substantial variability in spatial resolution and image quality. The number of tumors per case ranges from 1 to 6, and tumor sizes range from 4 mm to 68 mm in diameter. The test set contains a higher frequency of tumor occurrences than the training set. The average Hounsfield Unit (HU) value of tumors is 57.3 in the training set and 78.2 in the test set. Dataset statistics are summarized in Table 1. The training and test sets originate from different data sources. Therefore, PanTS allows thorough evaluation of AI generalization to unseen centers.

### 3.2 Dataset Contributors

The CT scans for the PanTS dataset come from 145 centers across 20 countries. As summarized in Figure 2, the CT scans from the training set are assembled from 13 publicly available abdominal CT datasets; the test set includes scans that are collected from 3 centers—University of California, San Francisco (UCSF), Polish Hospitals (PH), and Peking University Third Hospital (PUTH)—as well as the RSNA Abdominal Traumatic Injury CT (RATIC) dataset [54], which spans 23 centers across 14 countries. All data are anonymized, and the CT scans have been reviewed visually to preclude the presence of personal identifiers. The only processing applied to the CT scans is a transformation into a unified NIfTI format using NiBabel in Python. All CT scans from the training set can be downloaded from their official websites; ethics approval was not required. The use of test set has received IRB approval from Johns Hopkins Medicine under IRB00403268.

### 3.3 Annotation Protocol

The pancreatic tumors in the PanTS dataset were manually annotated by a team of 23 medical annotators with varying levels of expertise in pancreatic imaging, as summarized in Table 2. Each CT scan was annotated slice-by-slice using the MONAI-Label software [12, 19], with annotators assigning one of the pre-defined anatomical labels or marking the region as *Background* if it did not correspond to any defined structure. Initial tumor annotations were performed by annotators with $\geq$3 years of radiology experience. Each annotation was then reviewed by three additional annotators

[4] Benchmark results for this split are available at https://github.com/MrGiovanni/PanTS.

Table 1: **Characteristics of the PanTS dataset.** The PanTS training and test sets differ significantly across most clinical and imaging variables, including age, sex distribution, image resolution, and contrast phases. $p$-values were computed with the Mann–Whitney $U$ test. Notably, the test set contains a similar proportion of tumor cases but includes more non-contrast scans, making it a more challenging and realistic out-of-distribution benchmark. Tumor burden and pancreas size also vary between sets, reinforcing the need for robust generalization in model evaluation. These differences justify our dataset split design for assessing model performance under distributional shifts.

| Variable | Training set ($n$ = 9,901) | Test set ($n$ = 26,489) | $p$-value |
|---|---|---|---|
| Age, mean (SD) | 60.6 (13.0) | 58.5 (17.0) | $1.78 \times 10^{-7}$ |
| Sex | | | $7.87 \times 10^{-27}$ |
| Female, no. (%) | 2,358 (23.8) | 13,090 (49.4) | |
| Male, no. (%) | 2,923 (29.5) | 11,714 (44.2) | |
| Unknown, no. (%) | 4,620 (46.7) | 1,685 (6.4) | |
| In-plane spacing, mm (IQR) | 0.81 (0.74, 0.98) | 0.75 (0.70, 0.83) | 0.00 |
| Slice thickness, mm (IQR) | 1.25 (0.80, 2.50) | 1.25 (1.25, 2.50) | $5.13 \times 10^{-169}$ |
| Contrast phase | | | 0.00 |
| Non-contrast, no. (%) | 4,488 (45.3) | 3,920 (14.8) | |
| Portal venous, no. (%) | 2,895 (29.2) | 20,296 (76.6) | |
| Arterial, no. (%) | 2,450 (24.7) | 2,273 (8.6) | |
| Delayed, no. (%) | 68 (0.8) | 0 (0.0) | |
| Pancreatic tumor | | | |
| Yes, no. (%) | 1077 (10.9) | 2,829 (10.7) | |
| No, no. (%) | 8,824 (89.1) | 23,660 (89.3) | |
| Dilated duct | | | |
| Yes, no. (%) | 3,387 (34.2) | 11,180 (42.2) | |
| No, no. (%) | 6,514 (65.8) | 15,309 (57.8) | |
| Tumors per positive CT, no. (IQR) | 1.00 (1.00, 1.00) | 1.00 (1.00, 2.00) | $1.48 \times 10^{-65}$ |
| Tumor volume, $mm^3$ (IQR) | 4,749 (1,658, 11,479) | 12,667 (3,347, 32,238) | $4.07 \times 10^{-53}$ |
| Tumor HU value, mean (SD) | 57.3 (30.7) | 78.2 (59.0) | $1.54 \times 10^{-10}$ |
| Pancreas volume, $mm^3$ (IQR) | 74,669 (52,806, 95,892) | 74,480 (56,676, 92,892) | $8.75 \times 10^{-2}$ |
| Pancreas HU value, mean (SD) | 56.8 (36.4) | 85.6 (54.8) | 0.00 |

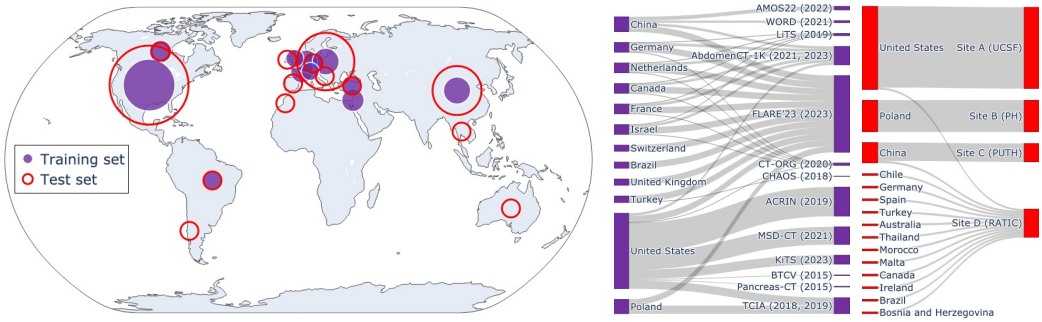


Figure 2: **Geographic diversity of the PanTS dataset.** Global distribution of contributing centers in the PanTS training set (purple circles) and test set (red outlines). Circle size is proportional to the base-10 logarithm ($\log_{10}$) of the number of CT scans contributed per country. The training set is aggregated from diverse public datasets spanning multiple countries, while the much larger test set is exclusively drawn from three independent centers—UCSF (United States, North America), PH (Poland, Europe), and PUTH (China, Asia)—not seen during training, as well as the RATIC dataset, which contributes scans from eight additional countries. This global coverage supports rigorous cross-institutional and out-of-distribution evaluation.

who were blinded to the initial labels. In cases of disagreement, a specialist served as the final arbiter to resolve labeling conflicts. Extremely small or ambiguous lesion-like structures were excluded to ensure consistency and quality. This structured multi-annotator annotation process was designed to ensure consistency, resolve ambiguity, and achieve high-quality voxel-wise annotations.

The PanTS dataset includes public organ and tumor segmentation datasets (Figure 2). However, these datasets were not fully-annotated for all tumors and structures we have in PanTS. The public datasets inside the PanTS training set had 191 pancreatic tumor annotations. We annotated 886 additional pancreatic tumors, reaching 1,077 pancreatic tumor annotations in the PanTS training set. Appendix A compares the number of structure annotations in public datasets and in PanTS. To

Table 2: **Annotator experience.** The 23 medical annotators contributing to the PanTS dataset span a wide range of experience levels, with Specialists averaging 27 years of practice, General radiologists 10 years, and Residents 4 years. Despite this variation, the annotators interpret a high volume of CT scans annually—Specialists averaging ∼10,300/year, Generals ∼18,000/year, and Residents ∼16,000/year—ensuring both breadth and depth of radiological expertise across annotations. This mix of senior and junior readers supports consistent, high-quality labeling while enabling scalability across thousands of cases.

| No. | Annotator ID | Experience (yr) | CT read / year | No. | Annotator ID | Experience (yr) | CT read / year |
|---|---|---|---|---|---|---|---|
| 1 | Specialist 1 (S1) | 24 | 12,000 | 2 | Specialist 2 (S2) | 22 | 12,000 |
| 3 | Specialist 3 (S3) | 35 | 8,000 | 4 | Specialist 4 (S4) | 30 | 8,000 |
| 5 | Specialist 5 (S5) | 28 | 9,000 | 6 | Specialist 6 (S6) | 19 | 13,000 |
| 7 | Specialist 7 (S7) | 23 | 11,000 | 8 | General 1 (G1) | 12 | 18,000 |
| 9 | General 2 (G2) | 8 | 18,000 | 10 | General 3 (G3) | 9 | 18,000 |
| 11 | General 4 (G4) | 10 | 18,000 | 12 | General 5 (G5) | 8 | 18,000 |
| 13 | General 6 (G6) | 13 | 18,000 | 14 | General 7 (G7) | 11 | 18,000 |
| 15 | General 8 (G8) | 10 | 18,000 | 16 | General 9 (G9) | 10 | 18,000 |
| 17 | General 10 (G10) | 13 | 18,000 | 18 | General 11 (G11) | 10 | 18,000 |
| 19 | Resident 1 (R1) | 5 | 16,000 | 20 | Resident 2 (R2) | 3 | 16,000 |
| 21 | Resident 3 (R3) | 4 | 16,000 | 22 | Resident 4 (R4) | 5 | 16,000 |
| 23 | Resident 5 (R5) | 5 | 16,000 | | | | |

efficiently scale voxel-wise annotations across pancreas head, body, tail, and 24 other anatomical structures, we employed a human-in-the-loop workflow [51, 38, 69]. Specifically, an AI-based anatomy segmentator was used to generate initial organ annotations, which were then manually verified and corrected by radiologists. This AI-assisted workflow was used only for non-tumor structures; all pancreatic tumors were annotated and reviewed manually.

### 3.4 Annotation Standard

Tumor annotations include the entire pancreatic mass, incorporating both solid and cystic components as well as intralesional necrosis, while excluding adjacent organs, fat, and vasculature. The pancreatic parenchyma is annotated into head, body, and tail based on anatomical landmarks: the head includes the uncinate process, and extends up to the mesenteric vessels; the body-tail separation is set at about the midpoint between the mesenteric vessels and the end of the pancreas tail. Only glandular tissue is included, excluding surrounding fat, vessels, and the duodenum. The pancreatic duct is annotated as a low-attenuation tubular structure extending from the tail to the ampulla of Vater, including both the duct wall and lumen, but excluding adjacent parenchyma and vessels. Related abdominal vessels are annotated as follows: the celiac artery from its origin to its trifurcation; the superior mesenteric artery (SMA) from its aortic origin to the first major branch; the portal vein from the confluence with the splenic vein to its entry into the liver; and the splenic vein from the splenic hilum to its confluence with the portal vein. For all vessels, both lumen and wall are included, while surrounding fat, organs, and unrelated tissues are excluded. Annotation standards for other vessels, abdominal organs, thoracic structures, and skeletal landmarks are detailed in the Appendix C.

### 3.5 Annotation Quality Control

Large medical image datasets inevitably contain annotation imperfections, particularly in voxel-wise annotations. While such datasets remain highly valuable, their utility can be further enhanced by systematically assessing annotation reliability. To evaluate internal consistency and quality of voxel-wise annotations in our training set, we conducted an inter-annotator agreement study (Figure 3E).

Specifically, we randomly selected 300 CT scans from the training set and had them independently re-annotated by a second radiologist, blind to the initial annotation. We computed the Dice Similarity Coefficient (DSC) between the two annotations for each case as a measure of agreement (Figure 4A). The median inter-annotator agreement was DSC (%) = 86.1%, with an interquartile range (IQR) of 19.6%, indicating high consistency across annotators. However, a small number of cases showed low agreement (DSC $<$ 20%), often due to small or ambiguous lesions. To ensure the annotation quality, we define a minimum threshold of DSC = 20% and flag all such cases for review and possible correction by senior radiologists.

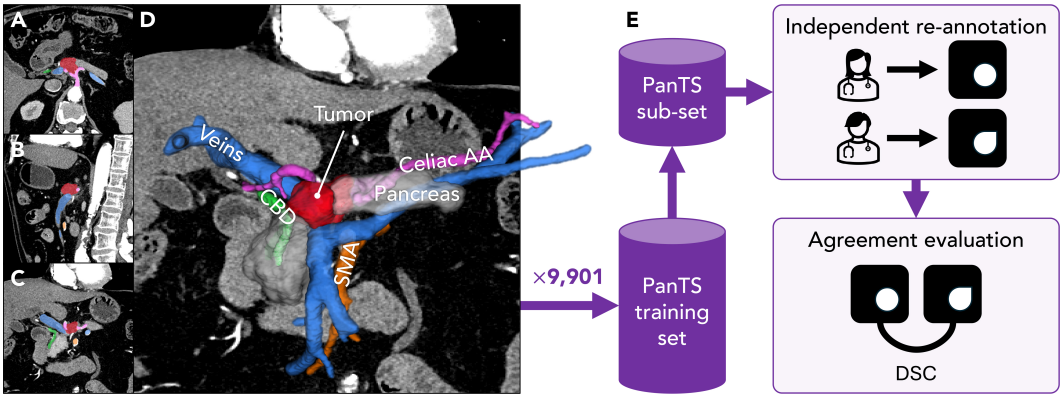


Figure 3: **Annotation standard and quality control. A–C.** Voxel-wise annotations of pancreatic tumors and surrounding anatomical structures shown on axial, sagittal, and coronal planes. Radiologists provide these annotations following the standard described in §3.4. **D.** 3D rendering on the coronal plane highlights detailed annotations of the tumor, pancreas, and key vessels, including the celiac artery (Celiac AA), superior mesenteric artery (SMA), common bile duct (CBD), and surrounding veins. **E.** To assess annotation quality, a subset of 300 CT scans from the PanTS training set was independently re-annotated by multiple radiologists. Inter-annotator agreement was evaluated using the Dice Similarity Coefficient (DSC).

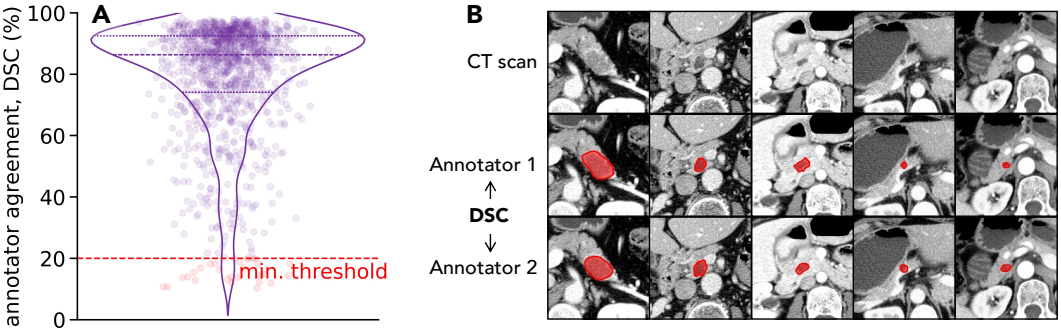


Figure 4: **Inter-annotator agreement on the PanTS subset. A.** Distribution of DSC (%) values between two independent radiologists across 300 CT scans from the PanTS training set. Most annotations demonstrate high agreement, confirming their reliability. A minimum threshold of DSC = 20% (dashed red line) is used to flag low-agreement cases, which are reviewed by senior radiologists for further quality assurance. **B.** Representative examples showing the same CT scan annotated by two different radiologists. High-agreement cases appear in the left columns, while low-agreement cases—often involving small or ambiguous lesions—appear on the right.

Figure 4B shows representative examples of CT scans annotated by two radiologists. High-agreement cases are shown on the left, while low-agreement cases—typically more subtle or ambiguous—are shown on the right. This inter-annotator evaluation not only ensures annotation quality control but also provides a reference for benchmarking automated models: systems that achieve DSCs comparable to or exceeding this agreement level can be considered human-comparable in segmentation performance.

## 4 Justification of Annotating Large-Scale Tumor Datasets

A central hypothesis is that scaling up voxel-wise tumor annotations significantly improves AI performance, particularly under out-of-distribution (OOD) settings—like hospitals not seen in training. To evaluate this, we trained a standard nnU-Net model on pancreatic tumor datasets of increasing size—MSD-Pancreas ($n$ = 281), PANORAMA ($n$ = 2,238), and our proposed PanTS dataset ($n$ = 9,901)—and evaluated detection performance on the held-out PanTS test set, which contains CT scans from medical centers not present in any training data.

As shown in Figure 5A, model performance improves with dataset scale, but not uniformly. The Area Under the ROC Curve (AUC) increases modestly from 0.810 (MSD) to 0.819 (PANORAMA), and then substantially to 0.959 when trained on our PanTS dataset[5]. While this trend partially aligns with AI scaling laws [30, 67]—which suggest that performance improves logarithmically with dataset size—the limited gain from MSD to PANORAMA indicates that scale alone is not sufficient. The significant improvement observed with PanTS is instead attributable to both its larger size and its high-quality, comprehensive annotations. PanTS includes 9,901 CT scans from 145 centers, capturing a broad range of pancreatic tumor types, anatomical variations, scan protocols, and noise distributions—factors essential for building robust, generalizable AI models.

To further assess the benefit of large-scale annotation, we benchmark nnU-Net trained on our PanTS dataset against leading AI models trained on MSD (Figure 5B). Using the official MSD test set, and third-party evaluated by the organizers of MSD challenge, our nnU-Net trained on PanTS outperforms all baseline methods by a margin of at least +4.9% DSC and +3.1% NSD in pancreatic tumor segmentation, becoming the new top-1 AI model in the public MSD-Pancreas leaderboard.

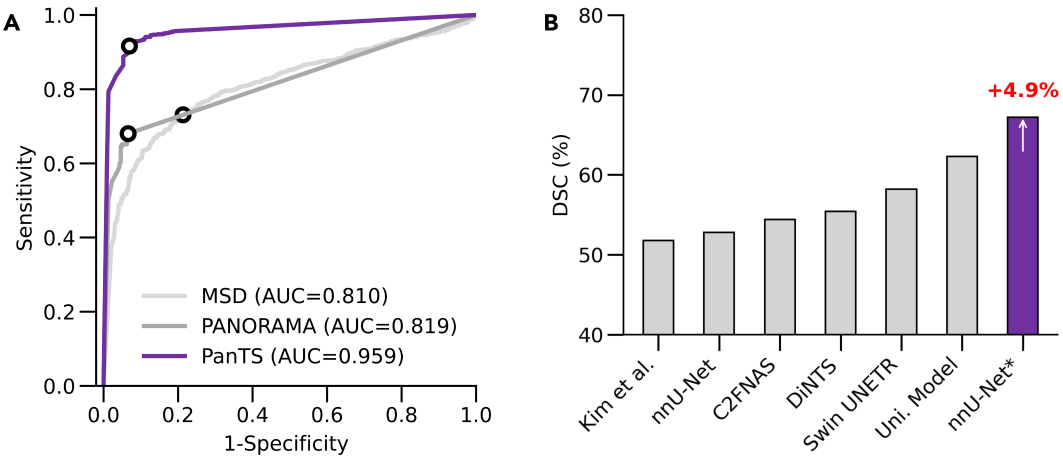


Figure 5: **Justification of annotating large-scale tumor datasets. A.** The Receiver Operating Characteristic (ROC) curve of standard nnU-Net trained on different scale of pancreatic CT datasets, i.e., MSD-Pancreas ($n$ = 281), PANORAMA ($n$ = 2,238), and our PanTS dataset ($n$ = 9,901). The performance is tested on the PanTS test dataset (CT collected different centers from MSD-Pancreas, PANORAMA, and the PanTS training set, detailed in Figure 2). The observation is the larger training set, the better pancreatic tumor detection performance on the out-of-distribution test set. **B.** Barplot of AI trained on our PanTS vs. AI trained on publicly available dataset (MSD-Pancreas). The performance is tested on the official MSD-Pancreas test set (third-party evaluation). All metrics can be found at The MSD Leaderboard.

## 5 Justification of Annotating 24 Surrounding Anatomical Structures

To assess the impact of anatomical context on pancreatic tumor segmentation, we compared the performance of a standard nnU-Net trained under two labeling schemes: a 2-class setup (tumor and pancreas) and a 28-class setup (tumor, pancreas subregions—head, body, tail—and 24 surrounding anatomical structures). Figure 6A shows the 28-class model markedly outperforms the 2-class model in tumor segmentation, with mean DSC improving **+10.3%** from 57.4% to 67.7%. Tumor boundary accuracy, measured by Normalized Surface Dice (NSD), also increases **+9.7%** from 56.8% to 66.5%.

By including structures such as the duodenum, bile duct, and nearby vessels, the 28-class model leverages additional spatial context to more effectively exclude non-tumorous tissue near ambiguous boundaries, enhancing spatial reasoning in anatomically complex regions. Annotating adjacent organs

[5]We hypothesize this discrepancy stems from annotation protocol differences: PANORAMA only annotates pancreatic ductal adenocarcinoma (PDAC), while treating all other tumors and healthy pancreases as *Normal*. This conflates distinct conditions under a single label, introducing ambiguity and limiting the model's ability to learn fine-grained distinctions between normal and abnormal tissue.

further encourages the model to internalize critical spatial relationships, especially in areas with low-contrast boundaries [29, 72]. These findings suggest that anatomical annotations function as implicit regularizers, helping the model structure its latent space more effectively.

The addition of 24 surrounding structures provides vital contextual cues, enabling clearer differentiation of tumors from neighboring tissues. This enriched anatomical supervision guides the model to learn spatial relationships, structural boundaries, and typical organ configurations—particularly important in the pancreas. These results highlight the importance of comprehensive multi-organ annotation for training robust and generalizable AI models in medical imaging.

In summary, our results confirm that including spatially related anatomical structures can improve segmentation of the class of interest. This underscores the importance of extensive anatomical annotation when designing large-scale, high-performance medical AI datasets.

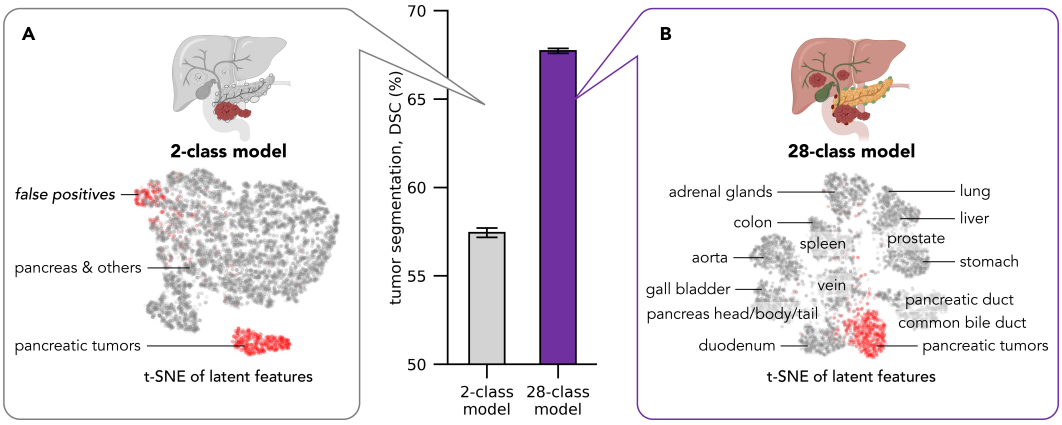


Figure 6: **Justification of annotating 24 surrounding anatomical structures.** We compare nnU-Net models trained with 2 classes (tumor and pancreas) versus 28 classes (tumor, pancreas head/body/tail, and 24 surrounding anatomical structures). The 28-class model significantly improves tumor segmentation accuracy (mean DSC +10.3%, $p < 0.0001$), highlighting the value of anatomical context. We further analyze the latent features of the two nnU-Net models. **A.** The 2-class model, trained to distinguish only pancreatic tumors vs. background, shows overlapping feature clusters in t-SNE space [63], with substantial false positives. **B.** The 28-class model, trained with supervision from 27 additional anatomical structures, results in better separation of pancreatic tumor features from surrounding tissues in t-SNE space.

## 6 Conclusion and Discussion

Our PanTS dataset marks a major advance in data-driven pancreatic cancer research. It includes more than 36,000 CT scans from 145 medical centers, enabling AI models that generalize across patient populations and imaging protocols. This dataset was built through a large collaborative effort involving 23 radiologists and years of annotation, quality control, and cross-validation. With nearly one million expert-validated voxel-wise annotations, PanTS is the largest public dataset for pancreatic tumor analysis to date.

We hope the release of PanTS will encourage more research groups to share medical datasets and annotations. We highlight two key aspects that we believe are especially important for public datasets in cancer related research.

***Normal CT scans matter.*** Public tumor datasets often include positive CT scans but contain few or no normal scans. For example, all the scans in MSD-Pancreas [6] contain pancreas tumors, so we won't know if AI trained on it is overly sensitive. No normal scan can be used to test it. Similarly, KiTS (for kidney tumors) [21], LiTS (for liver tumors) [10] datasets also offer a very limited number of normal scans. This imbalance makes it difficult to estimate the true negative rate (Specificity) and positive predictive value (PPV)[6]—two key metrics that determine whether an algorithm is suitable

[6]High PPV means the patient is very likely to have cancer if the AI predicts it.

for large-scale population screening. For example, in the general-population setting, where the prevalence of pancreatic tumor is extremely low, even a highly accurate model can yield many false positives. A simple Bayesian calculation illustrates the point: if 100,000 asymptomatic individuals are screened at 0.1% prevalence, even the state-of-the-art model (operating at 97% sensitivity and 99% specificity) would produce around 1,096 positive predictions, but only around 97 would be true positives (PPV = 8.9%). Most positive predictions in practice would be false, causing anxiety, overdiagnosis, and extra costs.

Our PanTS dataset helps address this evaluation gap by providing a large pool of normal CT scans (89% of both training and test sets), enabling assessment of number of false positives. We also provide both contrast-enhanced (*e.g.*, venous, arterial, delayed) and non-contrast CT scans, which enable opportunistic screening analyses in scans acquired not for cancer detection. The scale (36,390 CT scans from 145 centers) and rich labels allow model assessment beyond sensitivity alone and under clinically relevant operating points.

***Metadata matters.*** Because PPV depends on disease prevalence, screening will be more effective when focused on higher-risk groups rather than the general population. Integrating imaging biomarkers and clinical metadata (*e.g.*, age, contrast phase, ductal findings, notes) into a knowledge-graph or risk-score can raise effective prevalence and transform a population-level screener into a targeted detection tool. PanTS is designed for this: each scan includes metadata (age, sex, contrast phase, spacing, slice thickness), voxel-wise labels for the pancreas and 24 surrounding structures (*e.g.*, pancreatic duct, common bile duct, SMA, portal vein), and summary variables such as ductal dilatation—features that enable principled risk stratification and anatomy-aware modeling. In short, who we screen (risk stratification) and what we test on (abundant normal scans and diverse protocols) are as important as how we model. Datasets that pair many normal scans with rich metadata, as our PanTS does, are essential for developing models whose PPV and clinical value hold up in practice.

Despite its strengths, PanTS highlights the considerable challenges of annotating tumor datasets compared to normal anatomical structures. Even among experts, inter-annotator agreement can be modest, especially for small, ambiguous lesions. Our analysis of misclassified cases provides insight: in false positives, annotators noted subtle texture irregularities in the pancreas but without the hallmark signs of tumor presence (e.g., ductal dilation or parenchymal atrophy). Conversely, false negatives often involved subtle or atypical presentations, such as exophytic growths in hard-to-visualize regions as the pancreas tail or diffuse parenchymal thinning that may indicate underlying malignancy.

These findings underscore a central challenge: even experienced radiologists can miss early or atypical tumors, emphasizing the potential value of AI models trained on large, richly annotated datasets like PanTS. At the same time, they highlight the need for caution when interpreting both manual and automated annotations—especially in edge cases. Future work should explore multimodal learning, combining imaging, pathology, and clinical data, to further improve accuracy and reduce uncertainty.

Importantly, PanTS is more than a technical benchmark—it has clinical and translational significance. Pancreatic cancer remains one of the deadliest malignancies due to late-stage diagnoses and the subtlety of early radiologic signs. While AI holds promise for earlier detection, prior models have been hampered by small, homogeneous training data. By contrast, PanTS offers unprecedented scale and diversity, enabling the development of robust, generalizable AI systems. It also provides a foundation for anatomy-aware evaluation metrics, automated report generation, subpopulation analysis, and AI-assisted education. To maximize impact, we publicly release the baseline model and the PanTS training set under the non-commercial license.

## Acknowledgments and Disclosure of Funding

This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research, the Patrick J. McGovern Foundation Award, 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. We thank Jaimie Patterson for writing a news article about this project; thank Jaeden Pangaribuan, Zejun Wu, and Hsiang-Chen Yeh for developing the dataset website. We also thank Dariush Lotfi, affiliated with the Department of Diagnostic Radiology at The University of Hong Kong, for his valuable feedback and contributions to improving the PanTS dataset. Paper content is covered by patents pending.

## References


[1] Lorraine Abel, Jakob Wasserthal, Thomas Weikert, Alexander W. Sauter, Ivan Nesic, Marko Obradovic, Shan Yang, Sebastian Manneck, Carl Glessgen, Johanna M. Ospel, Bram Stieltjes, Daniel T. Boll, and Björn Friebe. Automated detection of pancreatic cystic lesions on ct using deep learning. *Diagnostics*, 11(5), 2021. ISSN 2075-4418. doi: 10.3390/diagnostics11050901. URL https://www.mdpi.com/2075-4418/11/5/901.

[2] Hugo J. W. L. Aerts, Emmanuel R. Velazquez, Ralph T. H. Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, Ronald Monshouwer, Benjamin Haibe-Kains, David Rietveld, Frank Hoebers, Marieke M. Rietbergen, René Leemans, Andre Dekker, John Quackenbush, Robert Gillies, and Philippe Lambin. A whole-body FDG-PET/CT dataset with manually annotated tumor lesions. *Scientific Data*, 9(1):1–9, 2022. doi: 10.1038/s41597-022-01718-3. URL https://www.nature.com/articles/s41597-022-01718-3.

[3] Oguz Akin, Paula Elnajjar, Mark Heller, Rosemary Jarosz, Bradley J. Erickson, Sheri Kirk, Yu Lee, Marston W. Linehan, Rajan Gautam, Ramesh Vikram, Kelly M. Garcia, Charles Roche, Emanuele Bonaccio, and Jeffrey Filippini. The Cancer Genome Atlas Kidney Renal Clear Cell Carcinoma Collection (TCGA-KIRC) (Version 3), 2016. URL https://doi.org/10.7937/K9/TCIA.2016.V6PBVTDR. [Data set].

[4] N Alves, M Schuurmans, D Rutkowski, et al. The panorama study protocol: Pancreatic cancer diagnosis-radiologists meet ai. zenodo, 2024.

[5] 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.

[6] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):1–13, 2022.

[7] Pedro RAS Bassi, Wenxuan Li, Yucheng Tang, Fabian Isensee, Zifu Wang, Jieneng Chen, Yu-Cheng Chou, Yannick Kirchhoff, Maximilian Rokuss, Ziyan Huang, Jin Ye, Junjun He, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus H. Maier-Hein, Paul Jaeger, Yiwen Ye, Yutong Xie, Jianpeng Zhang, Ziyang Chen, Yong Xia, Zhaohu Xing, Lei Zhu, Yousef Sadegheih, Afshin Bozorgpour, Pratibha Kumari, Reza Azad, Dorit Merhof, Pengcheng Shi, Ting Ma, Yuxin Du, Fan Bai, Tiejun Huang, Bo Zhao, Haonan Wang, Xiaomeng Li, Hanxue Gu, Haoyu Dong, Jichen Yang, Maciej A. Mazurowski, Saumya Gupta, Linshan Wu, Jiaxin Zhuang, Hao Chen, Holger Roth, Daguang Xu, Matthew B. Blaschko, Sergio Decherchi, Andrea Cavalli, Alan L. Yuille, and Zongwei Zhou. Touchstone benchmark: Are we on the right way for evaluating ai algorithms for medical segmentation? *Conference on Neural Information Processing Systems*, 2024. URL https://github.com/MrGiovanni/Touchstone.

[8] Pedro RAS Bassi, Qilong Wu, Wenxuan Li, Sergio Decherchi, Andrea Cavalli, Alan Yuille, and Zongwei Zhou. Label critic: Design data before models. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2025. URL https://github.com/PedroRASB/LabelCritic.

[9] 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. URL https://github.com/MrGiovanni/RadGPT.

[10] 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.

[11] Kai Cao, Yingda Xia, Jiawen Yao, Xu Han, Lukas Lambert, Tingting Zhang, Wei Tang, Gang Jin, Hui Jiang, Xu Fang, et al. Large-scale pancreatic cancer detection via non-contrast ct and deep learning. *Nature medicine*, 29(12):3033–3043, 2023.

[12] M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

[13] Luohai Chen, Wei Wang, Kaizhou Jin, Bing Yuan, Huangying Tan, Jian Sun, Yu Guo, Yanji Luo, Shi-Ting Feng, Xianjun Yu, et al. Special issue "the advance of solid tumor research in china": Prediction of sunitinib efficacy using computed tomography in patients with pancreatic neuroendocrine tumors. *International Journal of Cancer*, 152(1):90–99, 2023.

[14] Qi Chen, Xinze Zhou, Chen Liu, Hao Chen, Wenxuan Li, Zekun Jiang, Ziyan Huang, Yuxuan Zhao, Dexin Yu, Junjun He, et al. Scaling tumor segmentation: Best lessons from real and synthetic data. *arXiv preprint arXiv:2510.14831*, 2025. URL https://github.com/BodyMaps/AbdomenAtlas2.0.

[15] Yu-Cheng Chou, Zongwei Zhou, and Alan Yuille. Embracing massive medical data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 24–35. Springer, 2024. URL https://github.com/MrGiovanni/OnlineLearning.

[16] Linda C Chu, Michael G Goggins, and Elliot K Fishman. Diagnosis and detection of pancreatic cancer. *The Cancer Journal*, 23(6):333–342, 2017.

[17] Hwe Hoon Chung, Kyung Sook Lim, and Joo Kyung Park. Clinical clues of pre-symptomatic pancreatic ductal adenocarcinoma prior to its diagnosis: a retrospective review of ct scans and laboratory tests. *Clinics and Practice*, 12(1):70–77, 2022.

[18] 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. https://kaggle.com/competitions/rsna-2023-abdominal-trauma-detection, 2023. Kaggle.

[19] Andres Diaz-Pinto, Sachidanand Alle, Vishwesh Nath, Yucheng Tang, Alvin Ihsani, Muhammad Asad, Fernando Pérez-García, Pritesh Mehta, Wenqi Li, Mona Flores, et al. Monai label: A framework for ai-assisted interactive labeling of 3d medical images. *Medical Image Analysis*, 95:103207, 2024.

[20] 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.

[21] 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.

[22] Nicholas Heller, Sean McSweeney, Matthew Thomas Peterson, Sarah Peterson, Jack Rickman, Bethany Stai, Resha Tejpaul, Makinna Oestreich, Paul Blake, Joel Rosenberg, et al. An international challenge to use artificial intelligence to define the state-of-the-art in kidney and kidney tumor segmentation in ct imaging., 2020.

[23] J Hong, M Reyngold, C Crane, J Cuaron, C Hajj, J Mann, M Zinovoy, E Yorke, E LoCastro, AP Apte, et al. Breath-hold ct and cone-beam ct images with expert manual organ-at-risk segmentations from radiation treatments of locally advanced pancreatic cancer [data set]. the cancer imaging archive. *The Cancer Imaging Archive https://doi. org/10.7937/TCIA. ESHQ-4D90*, 2021.

[24] Sanne A Hoogenboom, Candice W Bolan, Anthony Chuprin, Maria T Raimondo, Jeanin E van Hooft, Michael B Wallace, and Massimo Raimondo. Pancreatic steatosis on computed tomography is an early imaging feature of pre-diagnostic pancreatic cancer: A preliminary study in overweight patients. *Pancreatology*, 21(2):428–433, 2021.

[25] Bowen Huang, Haoran Huang, Shuting Zhang, Dingyue Zhang, Qingya Shi, Jianzhou Liu, and Junchao Guo. Artificial intelligence in pancreatic cancer. *Theranostics*, 12(16):6931, 2022.

[26] 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.

[27] 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. *arXiv preprint arXiv:2404.09556*, 2024.

[28] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35:36722–36732, 2022.

[29] 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. URL https://github.com/MrGiovanni/LabelAssemble.

[30] 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.

[31] A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021.

[32] 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.

[33] Yoshihiro Konno, Yasuhiro Sugai, Masafumi Kanoto, Keisuke Suzuki, Toshitada Hiraka, Yuki Toyoguchi, and Kazuho Niino. A retrospective preliminary study of intrapancreatic late enhancement as a noteworthy imaging finding in the early stages of pancreatic adenocarcinoma. *European Radiology*, 33(7):5131–5141, 2023.

[34] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T 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*, volume 5, page 12, 2015.

[35] 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. URL https://github.com/MrGiovanni/SyntheticTumors.

[36] 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. URL https://github.com/MrGiovanni/AbdomenAtlas.

[37] Wenxuan Li, Alan Yuille, and Zongwei Zhou. How well do supervised models transfer to 3d image segmentation? In *International Conference on Learning Representations*, 2024. URL https://github.com/MrGiovanni/SuPreM.

[38] 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. URL https://github.com/MrGiovanni/ScaleMAI.

[39] 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. URL https://github.com/MrGiovanni/TextoMorph.

[40] 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. URL https://github.com/ljwztc/CLIP-Driven-Universal-Model.

[41] 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. URL https://github.com/ljwztc/CLIP-Driven-Universal-Model.

[42] 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.

[43] Xiangde Luo, Wenjun Liao, Jianghong Xiao, Jieneng Chen, Tao Song, Xiaofan Zhang, Kang Li, Dimitris N Metaxas, Guotai Wang, and Shaoting Zhang. Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. *Medical Image Analysis*, page 102642, 2022.

[44] Jun Ma, Yao Zhang, Song Gu, Yichi Zhang, Cheng Zhu, Qiyuan Wang, Xin Liu, Xingle An, Cheng Ge, Shucheng Cao, et al. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *arXiv preprint arXiv:2010.14808*, 2020.

[45] 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.

[46] Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Ershuai Wang, Qin Zhou, Ziyan Huang, Pengju Lyu, Jian He, and Bo Wang. Automatic organ and pan-cancer segmentation in abdomen ct: the flare 2023 challenge, 2024. URL https://arxiv.org/abs/2408.12534.

[47] Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Ershuai Wang, Qin Zhou, Ziyan Huang, Pengju Lyu, Jian He, and Bo Wang. Automatic organ and pan-cancer segmentation in abdomen ct: the flare 2023 challenge. *arXiv preprint arXiv:2408.12534*, 2024.

[48] Andrew McGuigan, Paul Kelly, Richard C Turkington, Claire Jones, Helen G Coleman, and R Stephen McCain. Pancreatic cancer: A review of clinical diagnosis, epidemiology, treatment and outcomes. *World journal of gastroenterology*, 24(43):4846, 2018.

[49] Ahmed W. Moawad, Diana Fuentes, Ahmed Morshid, Ahmed M. Khalaf, Mohamed M. Elmohr, Ahmed Abusaif, John D. Hazle, Ahmed O. Kaseb, Mohamed Hassan, Amir Mahvash, Jan Szklaruk, Ali Qayyom, and Khaled Elsayes. Multimodality annotated HCC cases with and without advanced imaging segmentation, 2021. URL https://doi.org/10.7937/TCIA.5FNA-0924. [Data set].

[50] Sovanlal Mukherjee, Anurima Patra, Hala Khasawneh, Panagiotis Korfiatis, Naveen Rajamohan, Garima Suman, Shounak Majumder, Ananya Panda, Matthew P Johnson, Nicholas B Larson, et al. Radiomics-based machine-learning models can detect pancreatic cancer on prediagnostic computed tomography scans at a substantial lead time before clinical diagnosis. *Gastroenterology*, 163(5):1435–1446, 2022.

[51] Chongyu Qu, Tiezheng Zhang, Hualin Qiao, Jie Liu, Yucheng Tang, Alan Yuille, and Zongwei Zhou. Abdomenatlas-8k: Annotating 8,000 abdominal ct volumes for multi-organ segmentation in three weeks. In *Conference on Neural Information Processing Systems*, volume 21, 2023. URL https://github.com/MrGiovanni/AbdomenAtlas.

[52] 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.

[53] 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.

[54] Jeffrey D Rudie, Hui-Ming Lin, Robyn L Ball, Sabeena Jalal, Luciano M Prevedello, Savvas Nicolaou, Brett S Marinelli, Adam E Flanders, Kirti Magudia, George Shih, et al. The rsna abdominal traumatic injury ct (ratic) dataset. *Radiology: Artificial Intelligence*, 6(6):e240101, 2024.

[55] Rebecca L Siegel, Angela N Giaquinto, and Ahmedin Jemal. Cancer statistics, 2024. *CA: a cancer journal for clinicians*, 74(1):12–49, 2024.

[56] Rebecca L Siegel, Tyler B Kratzer, Angela N Giaquinto, Hyuna Sung, and Ahmedin Jemal. Cancer statistics, 2025. *Ca*, 75(1):10, 2025.

[57] Dhruv Pratap Singh, Shannon Sheedy, Ajit H Goenka, Michael Wells, Nam Ju Lee, John Barlow, Ayush Sharma, Harika Kandlakunta, Shruti Chandra, Sushil Kumar Garg, et al. Computerized tomography scan in pre-diagnostic pancreatic ductal adenocarcinoma: Stages of progression and potential benefits of early intervention: A retrospective study. *Pancreatology*, 20(7):1495–1501, 2020.

[58] 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.

[59] Yucheng Tang, Jie Liu, Zongwei Zhou, Xin Yu, and Yuankai Huo. Efficient 3d representation learning for medical image analysis. *World Scientific Annual Review of Artificial Intelligence*, 2024.

[60] Tao Tong and Ming Li. Abdominal or pelvic enhanced CT images within 10 days before surgery of 230 patients with stage II colorectal cancer (StageII-Colorectal-CT), 2022. URL https://doi.org/10.7937/p5k5-tg43. [Dataset].

[61] Fumihito Toshima, Ryosuke Watanabe, Dai Inoue, Norihide Yoneda, Tatsuya Yamamoto, Naoki Sasahira, Takashi Sasaki, Masato Matsuyama, Kaori Minehiro, Ukihide Tateishi, et al. Ct abnormalities of the pancreas associated with the subsequent diagnosis of clinical stage i pancreatic ductal adenocarcinoma more than 1 year later: a case-control study. *American Journal of Roentgenology*, 217(6):1353–1364, 2021.

[62] 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.

[63] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[64] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5), 2023.

[65] 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.

[66] 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.

[67] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113, 2022.

[68] Lulu Zhang, Santosh Sanagapalli, and Alina Stoita. Challenges in diagnosis of pancreatic cancer. *World journal of gastroenterology*, 24(19):2047, 2018.

[69] Tiezheng Zhang, Xiaoxi Chen, Chongyu Qu, Alan Yuille, and Zongwei Zhou. Leveraging ai predicted and expert revised annotations in interactive segmentation: Continual tuning or full training? In *IEEE International Symposium on Biomedical Imaging*. IEEE, 2024. URL https://github.com/MrGiovanni/ContinualLearning.

[70] Yixiao Zhang, Xinyi Li, Huimiao Chen, Alan L Yuille, Yaoyao Liu, and Zongwei Zhou. Continual learning for abdominal multi-organ and tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 35–45. Springer, 2023. URL https://github.com/MrGiovanni/ContinualLearning.

[71] ZhiYu Zhao and Wei Liu. Pancreatic cancer: a review of risk factors, diagnosis, and treatment. *Technology in cancer research & treatment*, 19:1533033820962117, 2020.

[72] Zengle Zhu, Mintong Kang, Alan Yuille, and Zongwei Zhou. Assembling existing labels from public datasets to diagnose novel diseases: Covid-19 in late 2019. *NeurIPS Workshop on Medical Imaging meets NeurIPS*, 2022. URL https://github.com/MrGiovanni/LabelAssemble.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions by summarizing the dataset's unprecedented scale, rich anatomical annotations, and the demonstrated performance gains in tumor detection and segmentation, as evidenced by comprehensive experiments and ablation studies.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section 2 and Section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper does not present any theoretical results. Its contributions lie in the construction of a large-scale pancreatic CT dataset and the empirical validation of its effectiveness, rather than in formal assumptions or mathematical proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See Section 4, Section 5 and Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code, Models & Data:
https://huggingface.co/datasets/MrGiovanni/_PanTSMini

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: See Section 4, Section 5, and Appendix B.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: See Figure 4, Figure 6 and Appendix D.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Appendix E.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: This research fully adheres to the NeurIPS Code of Ethics. All patient-identifiable information was anonymized during preprocessing to ensure privacy protection, and the released dataset contains no identifiable information.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Appendix F.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The PanTS dataset poses no high risk for misuse. It consists of fully anonymized medical CT scans from clinical institutions and does not include any personally identifiable information, internet-scraped content, or generative models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: See Figure 1 for the use of existing datasets; Figure 5–6 for the use of existing code and models. A more detailed description is given in Appendix A, E, B.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have publicly released the training and evaluation code used in our benchmark (given in the abstract) and provided the download link of our datasets, i.e., PanTS.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects. All data were collected retrospectively from existing clinical records at participating institutions, and patient information was fully anonymized in compliance with ethical standards.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The study does not involve direct interaction with human subjects. All CT scans were retrospectively collected and fully anonymized prior to analysis.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core methodology of this research does not involve the use of LLMs in any important, original, or non-standard way. LLMs were not used in data processing, model development, or experimental evaluation.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

# Appendix

## Table of Contents

# A Related Datasets & Our Contribution

Table 3: **Comparison of PanTS with public abdominal CT datasets.** This comparative summary underscores the breadth, depth, and clinical relevance of PanTS relative to existing public datasets. While a number of prior datasets were incorporated into our training partition, our team made substantial and transformative contributions. Specifically, 23 board-certified radiologists independently annotated and rigorously validated previously unlabeled pancreatic tumors as well as over 25 additional abdominal and thoracic anatomical structures, many of which were not comprehensively labeled in the source datasets. This effort significantly elevates the clinical utility and completeness of the dataset. **Scale:** With 36,390 CT scans, PanTS is over 8.5× larger than the most extensive existing dataset dedicated to pancreatic tumor detection, setting a new benchmark for scale in abdominal imaging datasets. **Quality:** All tumor annotations meet silver-standard criteria, with expert oversight ensuring high inter-rater reliability and consistency. **Diversity:** The scans were collected from 145 institutions spanning 20 countries, offering a level of demographic and scanner variability that is 3× more diverse than previous benchmarks—critical for training generalizable and robust AI models. Collectively, these attributes make PanTS one of the most comprehensive, diverse, and clinically curated resources available for abdominal imaging research. *To advance transparency, reproducibility, and real-world relevance, we will publicly release the PanTS training set and use the PanTS test set to benchmark the performance of AI algorithms.*

| dataset | pancreatic tumors | number of CTs | institutions | countries | pancreas | pancr. head/body/tail | SMA | pancreatic duct | celiac artery | CBD | pancreatic veins | aorta | gallbladder | L kidney | R kidney | liver | IVC | spleen | stomach | L adrenal | R adrenal | bladder | colon | duodenum | L femur | R femur | L lung | R lung | prostate |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| KiTS'23 [2020] | 0 | 489 | 1 | 1 | × | × | × | × | × | × | × | × | × | ✓ | ✓ | × | × | × | × | × | × | × | × | × | × | × | × | × | × |
| LiTS [2019] | 0 | 131 | 7 | 5 | × | × | × | × | × | × | × | × | × | × | × | ✓ | × | × | × | × | × | × | × | × | × | × | × | × | × |
| TCIA-Pancr.-CT [2015] | 0 | 42 | 1 | 1 | ✓ | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × |
| CT-ORG [2020] | 0 | 140 | 8 | 6 | × | × | × | × | × | × | × | × | × | ✓ | ✓ | ✓ | × | × | × | × | × | ✓ | × | × | ✓ | ✓ | ✓ | ✓ | × |
| Trauma Det. [2023] | 0 | 4,714 | 23 | 13 | × | × | × | × | × | × | × | × | × | ✓ | ✓ | ✓ | × | ✓ | ✓ | × | × | × | ✓ | ✓ | × | × | × | × | × |
| BTCV [2015] | 0 | 47 | 1 | 1 | ✓ | × | × | × | × | × | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | × | × | × | × | × | × | × | × |
| CHAOS [2018] | 0 | 20 | 1 | 1 | × | × | × | × | × | × | × | × | × | × | × | ✓ | × | × | × | × | × | × | × | × | × | × | × | × | × |
| AbdomenCT-1K [2021] | 0 | 1,050 | 12 | 7 | ✓ | × | × | × | × | × | × | × | × | ✓ | ✓ | ✓ | × | ✓ | × | × | × | × | × | × | × | × | × | × | × |
| WORD [2021] | 0 | 120 | 1 | 1 | ✓ | × | × | × | × | × | × | × | ✓ | ✓ | ✓ | ✓ | × | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | × | × | × |
| AMOS [2022] | 0 | 200 | 2 | 1 | ✓ | × | × | × | × | × | × | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | × | ✓ | × | × | × | × | ✓ |
| MSD-CT [2021] | 191 | 945 | 1 | 1 | ✓ | × | × | × | × | × | × | × | × | × | × | ✓ | × | ✓ | × | × | × | × | ✓ | × | × | × | ✓ | ✓ | × |
| PANORAMA [2024] | 578 | 2,238 | 7 | 1 | ✓ | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × |
| **PanTS (ours)** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| training set | 1,076 | 9,901 | 119 | 12 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| test set | 2,829 | 26,489 | 26 | 15 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

# B Baseline and Implementation Details

## B.1 Top-Performing Methods in Medical Segmentation Decathlon

**Kim *et al.*** [32] proposed a neural architecture search (NAS) framework for 3D medical image segmentation tasks. This method explores a broad design space by automatically searching for optimal layer-wise structures, including both neural connectivities and operation types, across the encoder and decoder stages. To address the high computational cost associated with high-resolution 3D data, the framework introduces a scalable stochastic sampling algorithm based on continuous relaxation, which enables efficient gradient-based optimization.

**nnU-Net** [26, 27] is a self-configuring segmentation framework. It automatically configures pre-processing, network architecture, training, and post-processing. Its auto-configuration is guided by a combination of fixed parameters, interdependent rules that account for dataset characteristics and computational constraints, as well as empirical heuristics.

**C2FNAS** [66] is a coarse-to-fine neural architecture search (C2FNAS) framework designed to reduce the complexity and manual effort involved in developing task-specific 3D segmentation networks. This method addresses the common issue of inconsistency between the search and deployment stages in traditional NAS—often caused by memory limitations and expansive search spaces—by decoupling the architecture search into two successive phases. In the coarse stage, the framework explores the macro-level network topology, determining how convolutional modules are connected. In the fine stage, it refines the architecture by selecting specific operations within each cell, guided by the previously discovered topology. This coarse-to-fine strategy mitigates search-deployment mismatches while preserving scalability.

**DiNTS** [20] introduces a differentiable neural architecture search (NAS) framework tailored for 3D medical image segmentation, which aims to enable flexible topology design, high search efficiency, and controlled GPU memory usage. Unlike traditional NAS methods that are constrained by fixed topologies (e.g., U-Net) or suffer from long search times on large 3D datasets, DiNTS facilitates the automatic discovery of multi-path network topologies through a highly flexible and continuous search space. To address the discretization gap—the performance drop observed when converting an optimal continuous architecture into a discrete one—the method incorporates a topology loss to preserve the quality of the searched architecture. Furthermore, DiNTS integrates GPU memory constraints directly into the search process, making it more practical for resource-intensive 3D tasks.

**Swin UNETR** [58] adapted Swin Transformers to enhance medical image segmentation by capturing both local and global features through a hierarchical, window-based self-attention mechanism, outperforming the original UNETR by effectively modeling global context with Swin Transformers. Additionally, self-supervised pre-training of Swin Transformers on large-scale unlabeled 3D medical image datasets—using techniques such as masked autoencoding—can significantly boost model robustness and downstream task performance. These features led to state-of-the-art performance in various 3D medical image analysis applications, particularly in CT segmentation tasks.

**Universal Model** [40, 41, 70] was proposed to overcome the limitations of dataset-specific models in organ and tumor segmentation. Traditional models often suffer from poor generalizability due to the small size, partial annotations, and limited diversity of individual datasets. In contrast, the proposed model leverages text embeddings derived from Contrastive Language-Image Pre-training (CLIP) to encode anatomical labels. This enables the model to learn semantically structured feature representations and facilitates the segmentation of 25 organs and 6 tumor types across diverse anatomical regions. The model demonstrates strong transferability to novel domains and previously unseen tasks.

## B.2 Experimental Setting

### B.2.1 Justification of Annotating Large-Scale Tumor Datasets

To verify the effectiveness of scaling up voxel-wise tumor annotations and to justify the annotation of the PanTS dataset, we designed two comparative experiments to assess how increasing the volume of annotated data affects model performance, particularly in out-of-distribution (OOD) scenarios.

- Experiment 1: We selected two widely used public datasets—MSD-Pancreas ($n$ = 281) and PANORAMA ($n$ = 2,238)—as representative baselines for comparison with our proposed large-scale dataset, PanTS ($n$ = 9,901). A standard nnU-Net model was independently trained on each of the three datasets using identical configurations, including network architecture, data preprocessing, augmentation strategies, and optimization parameters, to ensure a fair comparison. All models were evaluated on the PanTS test set, which consists of CT scans from medical centers not included in the training data.
- Experiment 2: We benchmarked nnU-Net trained on the PanTS dataset against leading AI methods trained on the MSD dataset. Specifically, we selected Kim *et al.* , nnU-Net, C2FNAS, DiNTS, Swin UNETR, and Uni. Model as baselines for comparison, all trained on the MSD training set. The official MSD test set was used for evaluation, with performance independently evaluated by the organizers of the MSD challenge.

This experimental setting enables quantification of the benefits of large-scale tumor annotation by comparing model performance across datasets of increasing size and by evaluating under both in-distribution and out-of-distribution conditions.

#### B.2.2 Justification of Annotating 24 Surrounding Anatomical Structures

To evaluate whether incorporating detailed anatomical context improves the ability of tumor segmentation models to distinguish tumor boundaries, we conducted a comparative study under two labeling schemes. The core hypothesis is that segmenting additional surrounding structures enables the network to better capture anatomical boundaries and spatial relationships, thereby enhancing its ability to localize and delineate tumors.

Specifically, we trained the standard nnU-Net model using two distinct annotation protocols:

- A 2-class setup, including only the tumor and pancreas regions, reflecting the minimal annotation approach commonly used in public datasets.
- A 28-class setup, encompassing the tumor, pancreas subregions (head, body, and tail), and 24 surrounding anatomical structures, including vessels, gastrointestinal organs, and adjacent tissues.

Both models were trained on the same cohort of CT scans from the PanTS dataset, ensuring that performance differences are solely attributable to the inclusion of more comprehensive structural annotations. All training configurations—including preprocessing steps, augmentation strategies, and optimization parameters—were held constant across both setups. By comparing segmentation results on the held-out PanTS test set, we assessed whether finer-grained anatomical annotations enhance generalization performance and tumor localization accuracy.

### B.3 Implementation Details

#### B.3.1 Justification of Annotating Large-Scale Tumor Datasets.

- Experiment 1: The three standard nnU-Net models were trained using the nnU-Net framework. The orientation of CT scans was standardized to a consistent anatomical orientation. All preprocessing parameters—including resampling spacing, intensity range, and crop size—were automatically selected by the nnU-Net framework through empirical optimization on each training dataset. Detailed configuration settings are included in the accompanying code repository as JSON files. Data augmentation during training followed the default strategies defined by the nnU-Net framework. All models were trained for 1,000 epochs, each consisting of 250 iterations. We employed the SGD optimizer with a base learning rate of 0.01 and a batch size of 2. During inference, we applied test-time augmentation and used the sliding window strategy with an overlap ratio of 0.5, following the default nnU-Net implementations.
- Experiment 2: The training and inference procedures for our nnU-Net model followed the same configurations described in Experiment 1. For the comparative models, we report the official results released by the MSD Challenge organizers on the public leaderboard.

#### B.3.2 Justification of Annotating Large-Scale Tumor Datasets.

The two standard nnU-Net models were trained using the nnU-Net framework, following training procedures consistent with those described in Experiment 1. The only distinction between the two setups lies in the class labels used for training, with all other configurations kept identical.

### B.4 Evaluation Metrics

Each evaluation metric captures a specific aspect of the results, and selecting appropriate metrics is essential to highlight the characteristics of interest. To quantitatively evaluate segmentation performance, we employ a suite of widely adopted metrics: Dice Similarity Coefficient (DSC), Normalized Surface Dice (NSD), Sensitivity, Specificity, and Area Under the Receiver Operating Characteristic Curve (AUC).

#### B.4.1 Dice Similarity Coefficient (DSC)

DSC measures the volumetric overlap between the predicted segmentation and the ground truth. It is defined as:

$$\text{DSC} = \frac{2|P \cap G|}{|P| + |G|} \tag{1}$$

where $P$ and $G$ denote the sets of predicted and ground truth positive voxels, respectively. DSC ranges from 0 to 1, with higher values indicating better agreement. It is particularly useful for handling imbalanced data and is the standard metric in many medical imaging tasks.

#### B.4.2 Normalized Surface Dice (NSD)

NSD evaluates the agreement between the predicted and ground truth surfaces within a specified tolerance $\tau$, which reflects clinically acceptable deviation. It is defined as:

$$\text{NSD} = \frac{|\{x \in \partial P : \exists y \in \partial G, \|x - y\| < \tau\}| + |\{y \in \partial G : \exists x \in \partial P, \|y - x\| < \tau\}|}{|\partial P| + |\partial G|}, \tag{2}$$

where $\partial P$ and $\partial G$ represent the surfaces of the predicted and ground truth segmentations. NSD provides a more stringent surface-level evaluation, which is especially relevant in clinical applications requiring precise boundary delineation.

#### B.4.3 Sensitivity & Specificity

Sensitivity (also known as recall or true positive rate) quantifies the proportion of actual positives correctly identified, while Specificity measures the proportion of actual negatives correctly identified. They are defined as:

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \quad \text{Specificity} = \frac{TN}{TN + FP}, \tag{3}$$

where $TP$, $TN$, $FP$, and $FN$ are the numbers of true positives, true negatives, false positives, and false negatives, respectively. High sensitivity is critical for minimizing missed detections, whereas high specificity is important to reduce false alarms.

#### B.4.4 Area Under the Receiver Operating Characteristic Curve (AUC)

The AUC quantifies the overall ability of a model to discriminate between classes by measuring the area under the ROC curve, which illustrates the trade-off between sensitivity and specificity across varying thresholds. The ROC curve plots sensitivity against (1 – specificity) across different threshold values. An AUC value of 1.0 indicates perfect classification, while a value of 0.5 represents random guessing. AUC is particularly useful for evaluating the model's discriminative capability in segmentation tasks.

## C Annotation Standard

**Pancreas and Related Structures.** *Pancreatic tumors:* Annotate the entire tumor mass regardless of location within the pancreas. Include both solid and cystic components, as well as any intralesional necrosis. Exclude adjacent organs, fat, and vasculature. *Pancreas head, body, and tail:* Annotate the pancreatic parenchyma divided into three anatomical regions. The head is located to the right of the superior mesenteric vessels, within the curvature of the duodenum, and includes the uncinate process. The body lies between the left border of the superior mesenteric vessels and the left edge of the aorta. The tail lies anterior to the aorta, extending toward the splenic hilum. Include the entire gland parenchyma, excluding surrounding fat, vessels, and the duodenum. *Pancreatic duct:* Identify as a low-attenuation tubular structure within the pancreas. Annotate from the tail to the ampulla of Vater, including both the duct wall and lumen. Exclude surrounding pancreatic parenchyma and vessels.

**Vascular Structures.** *Aorta:* Annotate the entire lumen from the diaphragm to the bifurcation. Include the arterial wall and any calcifications, ulcers, thrombus, or dissection. Exclude surrounding tissues and organs. *Celiac artery:* Identify as a short arterial branch from the aorta. Annotate from its origin to its division into the left gastric, splenic, and common hepatic arteries. Include the lumen and wall. Exclude surrounding fat and organs. *Superior mesenteric artery (SMA):* Trace from its origin at the aorta to the point of major branching. Include the vessel wall and lumen. Exclude surrounding fat, pancreas, and bowel. *Postcava:* Annotate the entire lumen and wall from its origin at the postcava to its entry into the right atrium. Include any intraluminal thrombus. Exclude surrounding fat and structures. *Portal vein:* A bright, enhanced vessel formed by the confluence of the SMV and splenic vein. Annotate from the confluence to liver entry. Include lumen, wall, and any thrombus. *Splenic vein:* Trace from the spleen to its confluence with the SMV. Include lumen and wall, excluding adjacent pancreatic tissue and fat.

**Abdominal Organs.** *Liver:* Annotate the entire parenchyma including all segments, intrahepatic vessels, bile ducts, and any hepatic lesions. Exclude adjacent organs and fat. *Spleen:* Annotate the entire splenic parenchyma and any lesions. Exclude surrounding fat and nearby structures such as stomach, kidney, and colon. *Left and right kidneys:* Annotate the renal parenchyma. Exclude renal pelvis, ureter, perirenal fat, and adjacent structures. Include renal lesions if present. *Left and right adrenal glands:* Annotate the entire gland and any lesions. Exclude surrounding fat and nearby organs. *Gall bladder:* Annotate the wall and lumen, including the fundus, body, and neck. Include gallstones or polyps. Exclude cystic duct and liver tissue. *Stomach:* Annotate the entire wall and lumen including fundus, body, antrum, and pylorus. Include lesions. Exclude adjacent organs and fat. *Duodenum:* Annotate the wall and lumen from bulb to ligament of Treitz. Include lesions. Exclude pancreas, bile duct, and vasculature. *Common bile duct (CBD):* Identify as a low-attenuation tubular structure. Annotate from the hepatic duct confluence to the ampulla of Vater. Include duct wall and lumen. *Colon:* Annotate the wall and lumen of the cecum, appendix, ascending, transverse, descending, and sigmoid colon. Include lesions. Exclude fat, mesentery, and omentum. *Bladder:* Annotate the wall and lumen. Include intraluminal lesions. Exclude surrounding fat, muscles, and reproductive structures. *Prostate:* Annotate the entire parenchyma and prostatic urethra. Include lesions. Exclude surrounding fat, venous plexus, and seminal vesicles.

**Skeletal Structures.** *Left and right femurs (proximal):* Annotate the femoral head, neck, and up to 5 cm distal to the lesser trochanter. Include both cortical and cancellous bone and any lesions. Exclude surrounding muscles and vessels.

**Thoracic Organs.** *Left and right lungs:* Annotate the lung parenchyma, bronchovascular bundle, visceral pleura, and any lesions. Exclude pleural effusion, parietal pleura, mediastinal structures, and chest wall.

## D Additional Analysis of Benchmark Results

We participated in the Medical Segmentation Decathlon (MSD), a widely recognized benchmark designed to evaluate the generalizability and robustness of medical image segmentation algorithms across a diverse range of anatomical structures and imaging modalities. Among the ten segmentation tasks in the MSD, Task07 (pancreas and pancreatic tumor segmentation on portal venous phase CT) is especially challenging due to the pancreas's complex shape, small volume, and low-contrast tumors that are often hard to delineate from surrounding tissues.

Our method ranked first overall on Task07, achieving a Dice Similarity Coefficient (DSC) of 0.80 for pancreas segmentation and 0.52 for pancreatic tumor, outperforming all competing methods in both anatomical structure and lesion-level accuracy.

Compared to the original MSD winning entry by nnU-Net [27], which reported average DSCs of 0.69 for pancreas and 0.21 for tumor, our method improves segmentation accuracy by +11% and +31% respectively. This demonstrates the substantial impact of our pipeline in handling class imbalance, hard-to-segment tumors, and variable organ morphology.

Additionally, methods such as nnFormer, UNETR, and Swin UNETR, which leverage Transformer-based architectures, show modest improvements in pancreas segmentation (DSC around 0.74–0.76), but struggle in tumor segmentation (DSC consistently below 0.30). These models often underperform in capturing small or poorly contrasted tumors, likely due to their lack of task-specific supervision or fine-grained contextual priors.

# E Experiments Compute Resources

## E.1 Data Preprocess & Storage

To convert the raw CT volumes into the standardized format used in our experiments, we implemented a multi-step preprocessing pipeline that includes the following stages: (1) anonymization and DICOM to NifTi conversion; (2) CT intensity normalization by clipping Hounsfield Units (HU) to the range of -1000 to 1000, followed by reorienting all volumes to a consistent RPS (Right-Posterior-Superior) direction; (3) organ and lesion mask alignment; and (4) consolidation into structured multi-organ volumes. This pipeline was executed on a workstation equipped with a 64-core AMD Ryzen Threadripper 7980X CPU and 128 GB of RAM. No GPU acceleration was used during preprocessing. Parallelization across CPU threads allowed us to process 36390 CT volumes in under 90 hours. After preprocessing, the dataset containing volumetric CT images and per-voxel organ and tumor annotations across 28 anatomical regions required approximately 6.6 TB of storage. To ensure reproducibility and easy access, we structured the data according to standardized folder conventions and provided detailed metadata for each case.

## E.2 Model Training & Inference

All models were trained using a single NVIDIA RTX 4090 GPU with 24 GB of memory. The training process consumed approximately 8 GB of GPU memory and took approximately 18 hours to complete 1,000 epochs. During inference, the memory footprint was approximately 5 GB. Given the large size of the test set (26,489 CT scans), inference was performed in parallel across multiple GPUs to expedite evaluation. Specifically, we used a single server equipped with eight NVIDIA RTX 4090 GPUs, allowing the full test set to be processed in approximately two days.

## F Potential Negative Societal Impacts

PanTS provides a valuable and unprecedented resource for advancing pancreatic CT analysis; however, several potential societal risks must be acknowledged. First, large-scale datasets may inadvertently reinforce existing biases if the demographic or clinical distributions of the 145 participating centers do not adequately reflect the diversity of global patient populations. This can lead to models that exhibit reduced performance in underrepresented populations, thereby exacerbating healthcare disparities. Second, despite rigorous anonymization, the inclusion of detailed metadata (e.g., patient age, diagnosis, scan phase) raises privacy concerns, particularly in multi-institutional datasets containing rare conditions. Third, as models trained on PanTS demonstrate substantial performance improvements, there is a risk that such benchmarks may incentivize overfitting to dataset-specific anatomical or imaging characteristics, thereby limiting real-world generalizability. Finally, the growing availability and reliance on benchmark-driven evaluations may result in the misapplication or overreliance on AI systems in clinical workflows without sufficient regulatory oversight or clinical validation. These issues underscore the importance of ethical dataset curation, careful benchmark design, and responsible AI deployment in healthcare.