Appendix

Table of Contents

A Related Datasets & Our Contribution **25**

B Baseline and Implementation Details **26**
B.1 Top-Performing Methods in Medical Segmentation Decathlon 26
B.2 Experimental Setting 26
B.3 Implementation Details 27
B.4 Evaluation Metrics 28

C Annotation Standard **29**

D Additional Analysis of Benchmark Results **30**

E Experiments Compute Resources **31**
E.1 Data Preprocess & Storage 31
E.2 Model Training & Inference 31

F Potential Negative Societal Impacts **32**

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 [67] 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 [59] 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 [41, 42, 71] 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 τ, 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 ∂P and ∂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.