AbdomenAtlas-8K: Annotating 8,000 CT Volumes for Multi-Organ Segmentation in Three Weeks

Chongyu Qu[1]   Tiezheng Zhang[1]   Hualin Qiao[2]   Jie Liu[3]
Yucheng Tang[4]   Alan L. Yuille[1]   Zongwei Zhou[1,*]

[1]Johns Hopkins University   [2]Rutgers University
[3]City University of Hong Kong   [4]NVIDIA

Code & Data: https://github.com/MrGiovanni/AbdomenAtlas

Abstract

Annotating medical images, particularly for organ segmentation, is laborious and time-consuming. For example, annotating an abdominal organ requires an estimated rate of 30–60 minutes per CT volume based on the expertise of an annotator and the size, visibility, and complexity of the organ. Therefore, publicly available datasets for multi-organ segmentation are often limited in data size and organ diversity. This paper proposes an active learning procedure to expedite the annotation process for organ segmentation and creates the largest multi-organ dataset (by far) with the spleen, liver, kidneys, stomach, gallbladder, pancreas, aorta, and IVC annotated in **8,448** CT volumes, equating to **3.2 million** slices. The conventional annotation methods would take an experienced annotator up to 1,600 weeks (or roughly 30.8 years) to complete this task. In contrast, our annotation procedure has accomplished this task in three weeks (based on an 8-hour workday, five days a week) while maintaining a similar or even better annotation quality. This achievement is attributed to three unique properties of our method: (1) label bias reduction using multiple pre-trained segmentation models, (2) effective error detection in the model predictions, and (3) attention guidance for annotators to make corrections on the most salient errors. Furthermore, we summarize the taxonomy of common errors made by AI algorithms and annotators. This allows for continuous improvement of AI and annotations, significantly reducing the annotation costs required to create large-scale datasets for a wider variety of medical imaging tasks.

1   Introduction

Medical segmentation is a rapidly advancing task that plays a vital role in diagnosing, treating, and radiotherapy planning [36, 10, 31, 103, 39]. Building datasets of a substantial number of annotated medical images is critical for training and testing artificial intelligence (AI) models[1] [104, 51, 70]. However, medical datasets carefully annotated by an annotator are infeasible to create at a huge scale using conventional annotation methods [65] because performing per-voxel annotations is expensive and time-consuming [89, 54, 52]. As a result, publicly available datasets for multi-organ segmentation are often limited in data size (a few hundred) and organ diversity[2] as reviewed in Figure 1(a).

---

*Corresponding author: Zongwei Zhou (ZZHOU82@JH.EDU)

[1]While some of the abdominal organs can be automatically segmented by state-of-the-art AI models with fairly high accuracy, many other organs and anatomical structures remain suboptimal or unknown [54].

[2]TotalSegmentator [89] annotated a diverse set of classes for 1,024 CT volumes, but a majority of volumes were largely annotated *only* by pre-trained nnU-Nets [36]. As a result, their annotations are highly biased to the specific architecture (see §5); also, their procedure does not consider inconsistency across different AI models.
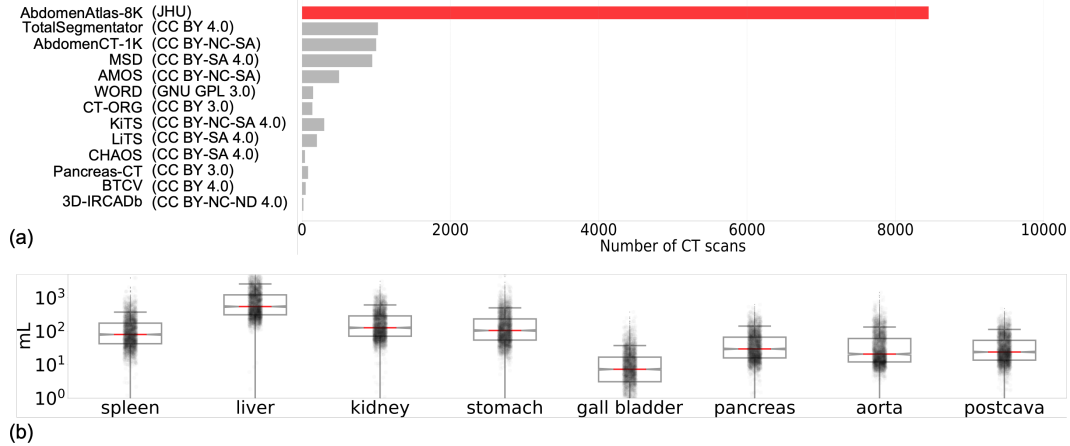
Figure 1: **(a) An overview of public datasets.** AbdomenAtlas-8K stands out from other datasets due to its large number of annotated CT volumes. We have reviewed dataset names (and licenses). **(b) Volume distribution of eight organs.** The significant variations within and across organs presented in our AbdomenAtlas-8K present challenges for the multi-organ segmentation and the generalizability of models to different domains. More comparisons can be found in Appendix Table 3 and Figure 9.

There is a pressing need to expedite the annotating procedure. The latest endeavors to construct a fully annotated dataset remained to ask radiologists to manually annotate each and every missing label [57, 35, 40, 78, 72, 9, 85, 20, 54, 99]—the same annotating procedure used a decade ago [19] or even earlier [64, 16]. Such a procedure is extremely costly, particularly for medical images, and will be time-consuming to create a large-scale dataset for every medical specialty [69, 90, 12, 83, 65]. To improve the annotation efficiency, active learning has been widely explored by combining the radiologist's competence and the computer's capability [106, 107, 71, 8, 11]. However, most studies in the active learning literature have been retrospective in nature, in which the annotating procedure was *simulated* by simply retrieving labels for the data without physically involving the radiologists in the loop to create/revise the labels [105, 66, 87, 94]. In contrast, our study is **proactive**, not only proposing a novel active learning procedure but also implementing it to actually construct a large dataset of 8,000 fully annotated CT volumes within a very short span of time, leveraging the synergy between medical professionals and AI algorithms in practice.

To overcome the deficiency of the conventional annotation method—which involves manually annotating each volume, slice, and voxel—we propose an efficient method to enable rapid organ annotation across enormous CT volumes. The efficiency has been demonstrated on 15 datasets and 8,448 abdominal CT volumes (1.2 TB in total). Our method enables high-quality, per-voxel annotations for eight organs and anatomical structures in all the CT volumes in three weeks (5 days per week; 8 hours per day), justified in §3.3. The constructed dataset is named AbdomenAtlas-8K, and its detailed statistics can be found in Figure 1. As a comparison, the conventional annotation methods [44, 25, 28, 50, 6] would take up to 1,600 weeks (30.8 years) to complete such a task[3].

Our key innovation is that rather than time-consuming annotating organs voxel by voxel, we leverage existing data and incomplete labels from an assembly of 16 public datasets. AI models are trained on the labeled part of data and generate predictions for a large number of unlabeled parts of data. We then actively revise the model predictions by only selecting the most salient part of the regions for annotators to correct. Moreover, our study provides a taxonomy of common errors made by AI algorithms and annotators. This taxonomy minimizes error duplication and augments data diversity, ensuring the sustainability of continuous revision of AI algorithms and organ annotations. Our novel active learning procedure involves only one trained annotator, three AI models, and commercial software (Pair[4]). After the procedure, annotators confirm the annotation of all the CT volumes by visual inspection. A large-scale, but private, dataset [92, 93] is used for external validation.

---

[3]Trained annotators annotate abdominal organs at a rate of 30–60 minutes per organ per three-dimensional CT volume reported by Park et al. [65]. We have also conducted a reader study to estimate the annotation time for each CT volume in our dataset, detailed in Appendix Table 4.

[4]We purchased a license from Pair. Alternatively, we also used open-source MONAI Label for annotation.

In summary, we made two major contributions. **Firstly**, AbdomenAtlas-8K was a composite dataset that unified datasets from 26 different hospitals worldwide. In total, over 60.6×10^9 voxels were annotated in comparison with 4.3×10^9 voxels annotated in the public datasets. We scaled up the organ annotation by a factor of 15 and released the masks for 5,195 of the 8,448 CT volumes. This large-scale, multi-center dataset can also impact downstream clinical tasks such as surgery, treatment, abdomen atlas, and anomaly detection. **Secondly**, the proposed active learning procedure can generate an attention map to highlight the regions to be revised by radiologists, reducing the annotation time from 30.8 years to three weeks and accelerating the annotation process by an impressive factor of 533. This strategy can quickly scale up annotations for creating many other medical imaging datasets.

2   Related Work

Large dataset construction. Kirillov et al. [42] created a huge natural image dataset of 1B masks and 11M images, but it lacks semantic information, and its efficacy is limited when applied to 3D volumetric medical images [51, 33]. Our AbdomenAtlas-8K, containing 8,448 CT volumes with per-voxel annotated eight abdominal organs, is the largest annotated CT dataset at the time this paper is written. We hereby review the existing public datasets that contained over 500 CT volumes with per-voxel annotated organs [22]. For example, AMOS [39], TotalSegmentator [89], and AbdomenCT-1K [54] provided 500, 1,024, and 1,112 annotated CT volumes, respectively. Both TotalSegmentator and AMOS derived their data from a single country, with the former reflecting the Central European population from Switzerland and the latter representing the East Asian population from China. In comparison, AbdomenAtlas-8K presented a greater data diversity because the CT volumes were collected and assembled from at least 26 different hospitals worldwide. While AbdomenCT-1K sourced data from 12 hospitals, AbdomenAtlas-8K contained approximately eight times the CT volumes (8,448 vs. 1,000) and twice the variety of annotated organs (8 vs. 4). Concurrently, we are actively expanding the classes covered by AbdomenAtlas-8K. Starting with the set of 104 classes found in TotalSegmentator, we aim to significantly diversify the range of classes covered.

Active learning for segmentation. *Uncertainty* and *diversity* are key criteria in active learning. Uncertainty-based criteria assess the value of annotating a data point based on the uncertainty (e.g., entropy) of AI predictions [17, 18, 55, 75, 5, 63, 11]. On the other hand, diversity-based criteria aim to select unannotated samples that differ from each other and from those already annotated [47, 24, 43, 58, 81, 82, 76]. For additional active learning methods, we refer the reader to comprehensive literature reviews [83, 61, 30, 71]; but these methods face computational complexity challenges with segmentation tasks and large unannotated data pools. To overcome this, we summarized typical errors made by humans and computers. Our active learning procedure considered the anatomical priors, uncertainty in AI prediction, and data diversity, and, importantly, pivoted a prospective application of active learning rather than retrospective studies. Moreover, the derived criteria can generate an attention map, pinpointing areas necessitating revision, thereby enabling precise detection of high-risk prediction errors (evidenced in Table 1). Consequently, our strategy markedly diminished the workload and annotation time for annotators by a factor of 533.

3   AbdomenAtlas-8K

Overview. We propose an active learning procedure comprising two components: error detection from AI predictions (§3.1) and manual revision performed by radiologists to review and edit the most significant errors detected (§3.2). By repeatedly implementing these two components, it is possible to expedite the creation of fully annotated datasets for multi-organ semantic segmentation. Finally, we describe the data construction strategies (§3.3). In this work, we have applied our approach to 8,448 CT volumes in portal (44%), arterial (37%), pre- (16%), and post-contrast (3%) phases.

3.1   Label Error Detection Revealed by Attention Maps

Figure 2 shows the process to generate attention maps for eight target organs. The attention maps can localize the potential error regions for human annotators to review and edit AI predictions.

(1) Inconsistency. To quantify inconsistency, we calculate the standard deviation of the soft predictions produced by multiple AI architectures, including Swin UNETR, nnU-Net, and U-Net. Regions

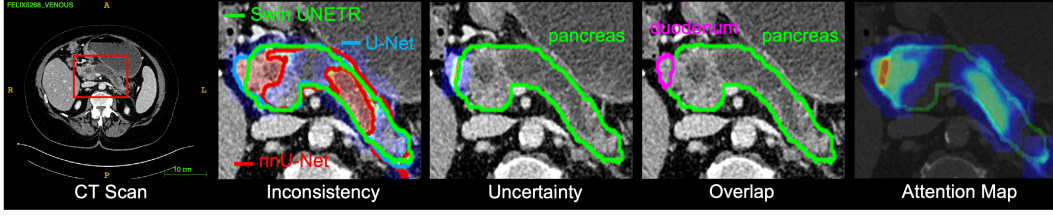| CT Scan | Inconsistency | Uncertainty | Overlap | Attention Map |

Figure 2: **Attention map generation.** The criteria inconsistency, uncertainty, and overlap refer to regions where model predictions diverge, exhibit high entropy values, and where multiple organ predictions overlap, respectively. The attention map visualizes a combination of these regions, drawing radiologists' attention to where AI predictions might falter. A standard color scheme helps in highlighting regions that merit closer review and revision. More examples are in Appendix Figure 11.

with high standard deviation indicate higher inconsistency and may require further revision.

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

where the subscript c represents class c of our eight target organs. For each voxel i, $p_{i,c}^n$ represents the soft prediction value obtained from the n-th AI architecture of class c at that voxel's index i, ranging from 0 to 1. $\mu_{i,c}$ represents the average prediction value obtained by combining the results of three AI architectures at the same voxel index. In this study, there are three AI architectures, so N is equal to three. Then, the Inconsistency$_{i,c}$ value is determined by the standard deviation of the soft prediction values from the three AI architectures.

(2) Uncertainty. To estimate the degree of certainty linked with the AI prediction of eight target organs, we determine the entropy of the soft predictions for each organ. Regions of higher entropy values suggest diminished confidence and increased ambiguity, potentially escalating the chances of encountering prediction errors within that specific area [98], which may necessitate further revision.

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

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

(3) Overlap. The overlap in organ prediction can indicate potential errors. If a voxel is predicted to be a part of both the liver and kidneys, even without ground truth, we can reasonably forecast a prediction mistake. We use the following measure to detect organ overlap in predictions.

$$\text{Overlap}_{i,c} = \begin{cases} 1 & \text{if } p_{i,c}^n > 0.5 \text{ and } \exists\, p_{i,c_{\not\in}}^n > 0.5 \\ 0 & \text{otherwise} \end{cases}$$ (3)

We generate *pseudo labels* by applying a threshold of 0.5 to the probability values. Pseudo labels refer to organ labels predicted by AI models without any additional revision or validation by human annotators. The overlap value, denoted as Overlap$_{i,c}$, is determined based on the following criteria: if the prediction value for class c exceeds the threshold of 0.5 for at least one AI architecture and there exists a prediction value not belonging to class c that exceeds 0.5 for the same voxel index i, then the overlap value is set to 1; otherwise, it is set to 0.

As a result, an *attention map* is generated to help annotators quickly locate regions that require revision or confirmation. We combine the inconsistency, uncertainty, and overlapping regions to produce the attention map. Consequently, a higher Attention$_{i,c}$ value in the 3D attention map indicates a greater risk of a prediction error for that voxel.

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

To assess the attention map, we identified error regions which is the summation of all false positive (FP) and false negative (FN) areas between the ground truth annotations (available in JHH [92]) and the pseudo labels predicted by AI. We then compare the attention maps with the error regions and calculate the sensitivity and precision, reported in §4.1.

4

3.2 Active Learning Procedure

Algorithm. Our active learning procedure has eight steps. ① Train an AI model from scratch, denoted as $\mathcal{M}_0$, using 2,100 CT volumes from 16 partially labeled public datasets. This takes approximately 40 hours. ② Direct test the current model (e.g., $\mathcal{M}_0$) on all 8,448 CT volumes to segment eight organs. This takes around 12 hours. ③ Compute organ-wise attention maps for each CT volume using the criteria of inconsistency, uncertainty, and overlap (§3.1), which highlight the regions that potentially have prediction errors and require human revision. ④ Compile a priority list sorted by the sum of attention maps. The larger the attention map is, the more urgent the CT volume requires revision. ⑤ Ask the annotators to review the top 5% (analyzed from Figure 4) CT volumes from the list and revise the pseudo labels guided by the attention maps. ⑥ Reassemble the annotation of each revised CT volume based on the label priority (detailed in the next paragraph). ⑦ Fine-tune the current model ($\mathcal{M}_t$) using the reassembled annotation to obtain $\mathcal{M}_{t+1}$. ⑧ Repeat steps ②-⑦ until the annotators confirm that the CT volume on the top of the prioritized list in step ④ does not need further revision, suggesting that AI predictions of the most important CT volumes have minimal errors.

Technical details in ⑥: label priority. In the assembly process, the *utmost priority* is given to the original annotations supplied by each public dataset. Subsequently, we assign *secondary priority* to the revised labels from our annotators. The pseudo labels, generated by AI models, are accorded the *lowest priority*. Based on this label priority, we reassemble the labels for each CT volume. This involves three different scenarios: inherited from the original labels of the public dataset if available, revised (if our annotator confirms that the pseudo labels have errors), and unchanged (if our annotator confirms that the pseudo labels predicted by the AI are correct). Figure 3 displays examples of the reassembled annotations from AbdomenAtlas-8K.
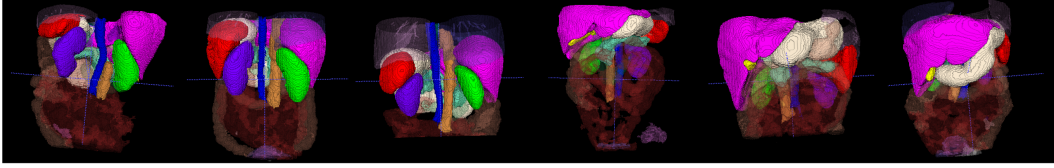

Figure 3: **AbdomenAtlas-8K annotations.** The annotations for our eight target organs (opaque) have been revised by our annotator or inherited from the original annotations of the partially labeled public datasets. The remaining organs and tumors (transparent) can be used for future research.
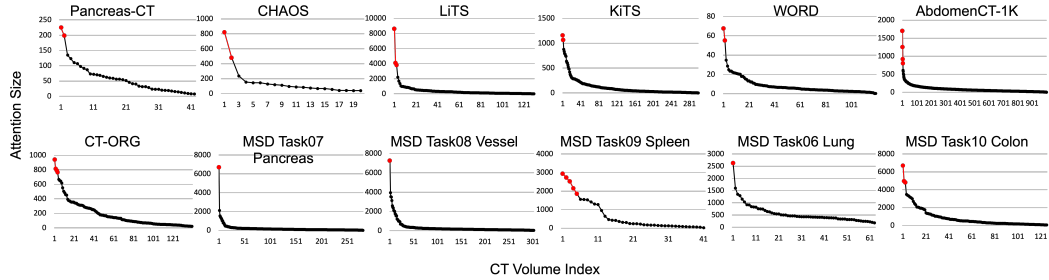

Figure 4: **Attention size distribution.** The y-axis denotes the attention size, the sum of Equation 4 over eight classes; each point corresponds to a distinct CT volume. A larger attention size implies a greater need for revision in various regions. While most CT volumes exhibit a small attention size, a few notable outliers (marked in red) stand out. These outliers are of high priority for revision by human experts. According to the figure, the ratio of outliers is about 5% (highlighted in red). The 5% is estimated by the plot and also related to the budget of human revision for each Step in the active learning procedure. It is essential to emphasize that roughly 5% of CT volumes within each dataset are highly likely to contain predicted errors, requiring further revision by our annotator.

Figure 4 illustrates the attention size of CT volumes within each partially labeled public dataset. It is important to note that the BTCV [45] and AMOS22 [39] datasets already have original annotations for our eight target organs. By analyzing the curve depicting the decreasing value of the attention map, we observed that among the 5,195 CT volumes from the 16 partially labeled abdominal datasets, only
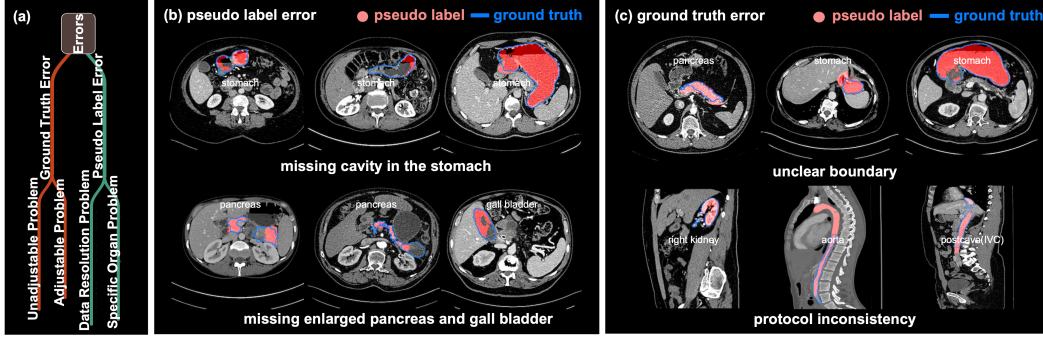
Figure 5: **(a) Error taxonomy.** To minimize repetitive errors during human revision, we collate and analyze common mistakes that occur during the data made by either AI or human annotators. **(b) Errors in pseudo labels.** For example, the cavity in the stomach is missing; the enlarged pancreas tail caused by a pancreatic tumor is missing; the enlarged gall bladder is missing. **(c) Errors in human annotations.** The high intra-annotator variability can be attributed to the ambiguity in defining organ or tumor boundaries, whereas the high inter-annotator variability is often a result of inconsistent annotation protocols and guidelines across different institutes.

5% of the volumes exhibited significantly large attention map values[5]. These high values indicate a higher risk of prediction errors, suggesting the need for further revision. Extrapolating this finding to a larger dataset of 8,000 CT volumes, we can estimate that the annotator will need to confirm and revise approximately 400 CT volumes. Assuming a rate of 15 minutes per CT volume and an 8-hour workday, this process would take approximately 12.5 days to complete. The attention map values are expected to decrease after fine-tuning $\mathcal{M}_0$ with revised labels as a new benchmark.

Figure 5(a) summarizes typical errors encountered in the active learning procedure into two categories: pseudo label errors and ground truth errors. As illustrated in Figure 5(b), pseudo label errors refer to the errors in the AI predictions. These errors often arise from irregular organ shapes, such as the absence of a cavity in the stomach or the omission of an enlarged pancreas and gall bladder. Discrepancies in CT volumes between the training and testing data, such as variations in scanners, protocols, reconstruction methods, and contrast enhancement, can also contribute to these inaccuracies. These issues result in the model's predictions lacking accuracy in representing these specific anatomical structures. As exampled in Figure 5(c), ground truth errors are errors in AI model predictions that result from inaccuracies in the human annotations used for training the model. These inaccuracies may arise due to unclear organ boundaries or inconsistency in labeling protocols across different institutions, introducing variations into the model's predictions.

3.3 Dataset Construction

Annotators. Our study recruited three annotators, comprising a senior radiologist with over 15 years of experience and two junior radiologists with three years of experience. The senior radiologist undertook the task of annotation revision in the active learning procedure. Before releasing AbdomenAtlas-8K, two junior radiologists looked through the masks in the entire AbdomenAtlas-8Kand made revisions if needed (i.e., our method missed the error regions)[6]. In addition, the two junior radiologists conducted the inter-annotator variability analysis (Figure 8) and recorded the time for conventional methods when each organ must be annotated voxel by voxel (Appendix Table 4).

Efficiency. *Why 30.8 years?* We considered an 8-hour workday, five days a week. A trained annotator typically needs 60 minutes per organ per CT volume [65]. Our AbdomenAtlas-8K has a total of eight organs and around 8,000 CT volumes. Therefore, annotating the entire dataset requires 60×8×8000 (minutes) / 60/8/5 = 1600 (weeks) = 30.8 (years). *Why three weeks?* Using our

---

[5]The 5% is empirically estimated based on (1) the observation in the distribution of the attention size (i.e., the number of outliers in Figure 4) and (2) the annotation budget at each Step in the active learning procedure. If there are many outliers or a limited budget, the threshold needs to be increased accordingly.

[6]Such revisions were seldom required based on our study—only 55 out of 8,448 volumes needing adjustments.

Table 1: **Evaluation metrics on JHH.** The sensitivity and precision are evaluated between our organ attention maps and organ error regions. Our attention map exhibits high average sensitivity and precision, indicating its effectiveness and accuracy in detecting false positives (FP) and false negatives (FN). This highlights its capability to precisely identify regions that need revision.

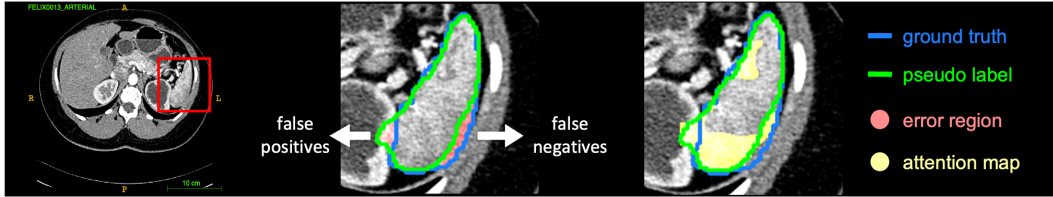| Metrics	Spl	RKid	LKid	Gall	Liv
Sensitivity	0.91 ± 0.25	0.93 ± 0.17	0.92 ± 0.18	0.99 ± 0.07	0.74 ± 0.33
Precision	0.68 ± 0.40	0.90 ± 0.22	0.85 ± 0.23	0.67 ± 0.33	0.88 ± 0.12
Metrics	Sto	Aor	IVC	Pan	**Avg.**
Sensitivity	0.98 ± 0.10	0.98 ± 0.11	0.98 ± 0.09	0.90 ± 0.21	0.93 ± 0.17
Precision	0.85 ± 0.16	0.82 ± 0.21	0.75 ± 0.22	0.91 ± 0.15	0.81 ± 0.22

Figure 6: **Error region vs. attention map.** The discrepancy between the model's pseudo labels and the ground truth delineates the error regions in predictions, comprising both false positives (FP) and false negatives (FN). These error regions serve as the benchmark for evaluating the sensitivity and precision of our attention map (§4.1).

active learning strategy, only 400 CT volumes require manual revision from the human annotator (15 minutes per volume). That is, we managed to accelerate the annotation process by a factor of $60 \times 8/15=32$ per CT volume. Therefore, we completed the entire annotation procedure within three weeks, as reported in the paper. Human efforts: 400×15 (minutes) / 60/8 = 12.5 (days) plus the time commitment for training and testing AI models taking approximately 8.5 (days).

Dataset splits. For 8,448 CT volumes in AbdomenAtlas-8K, we split them into training (5500 CT volumes), validation (500), and test (2448) sets. Each volume contains per-voxel annotations of the spleen, liver, left& right kidneys, stomach, gallbladder, pancreas, aorta, and IVC. Note that the JHH dataset is a proprietary, multi-resolution, multi-phase collected from Johns Hopkins Hospital [92, 41, 100, 46]. This dataset is used for external validation, including the assessment of the attention map quality (Table 1), AI generalizability (Table 2), the improvement of label quality in the active learning procedure (Appendix Table 5), and the performance of AI trained on a combination of public datasets and our AbdomenAtlas-8K(Appendix Table 6).

4 Experiment & Result

4.1 Attention Map and Annotation Evaluation

Evaluation of attention map. Our attention map is evaluated on 1,000 CT volumes of the JHH dataset [92], which is not included in the training process. The evaluation is performed across the eight target organs. Once an error is detected, it counts as a hit; otherwise, as a miss. To evaluate the quality of the attention map, two metrics are used: Sensitivity = TP / (TP + FN) and precision = TP / (TP + FP), where a true positive (TP) means the attention map found real mistakes the AI made; a false negative (FN) means it missed some of the AI's mistakes; and a false positive (FP) means it found mistakes where the AI was actually right. Sensitivity and precision are all calculated at the volume (a group of voxels) level rather than the voxel level. We chose sensitivity and precision because this experiment is designed to evaluate an error detection task rather than a segmentation task (comparing the boundary of attention maps and error regions). They can measure how well the attention maps detect the real error regions and whether the errors being detected are real errors, respectively. The results of the attention map evaluation are presented in Table 1. The mean values of sensitivity and precision metrics for our attention maps, with respect to eight target organs, are 0.93 ± 0.17 and 0.81 ± 0.22. Figure 6 shows the error regions and our attention map. These results
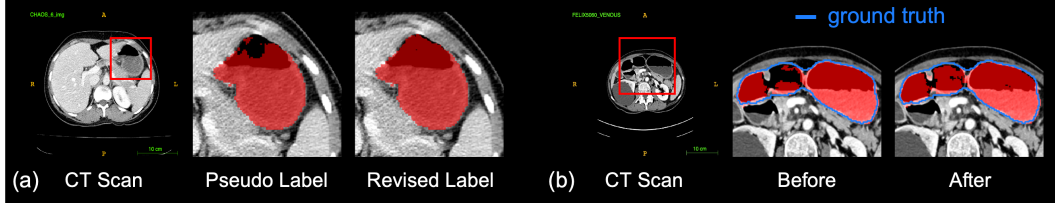
Figure 7: **(a) Pseudo vs. revised labels.** Pseudo labels are predicted by AI and then revised by annotators based on attention maps, resulting in revised labels. **(b) AI predictions before vs. after fine-tuning.** The post-fine-tuning results exhibit superior accuracy in segmenting the organ compared to the pre-fine-tuning result. Appendix Table 5 quantifies the DSC scores before and after fine-tuning.

(a) AI architecture variability
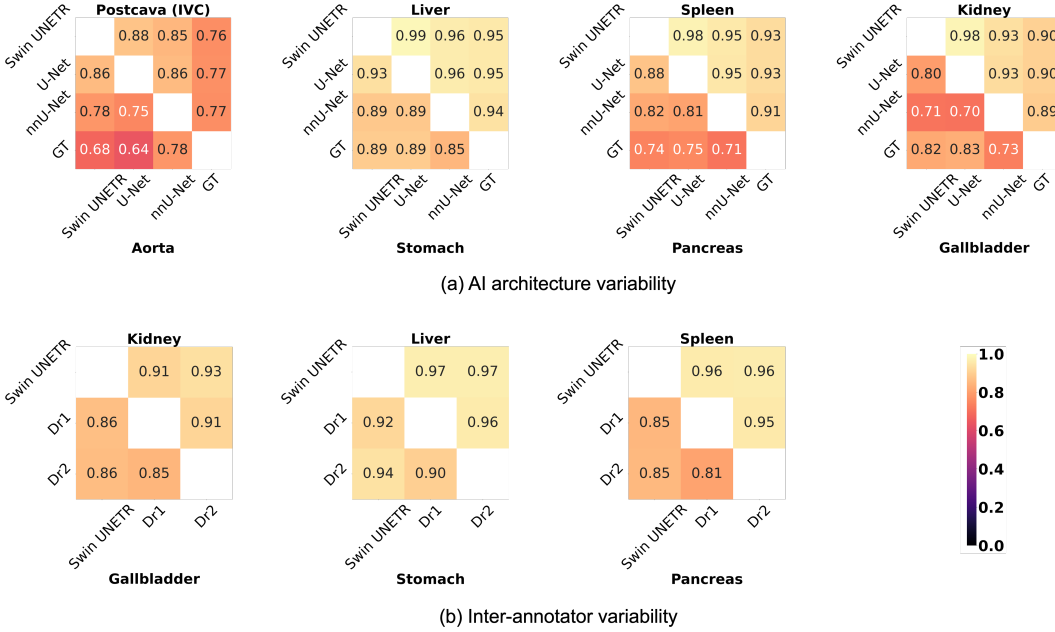

(b) Inter-annotator variability

Figure 8: The DSC score in each cell is calculated between the corresponding row and column. **(a) AI architecture variability.** The segmentation predictions made by different AI architectures display minor variations for eight organs, as evidenced by the corresponding DSC scores. Overall, the AI predictions align closely with the annotations provided by human experts. **(b) Inter-annotator variability.** The DSC scores between AI predictions and human annotators consistently outperform the score between the two human annotators, suggesting a comparable annotation quality between AI predictions and human annotations in the segmentation of six organs.

suggest that our attention map effectively captures the false positives and false negatives in error regions, demonstrating a high degree of accuracy in identifying prediction errors.

Evaluation of human annotation quality. We provided the visualization of pseudo labels (generated by AI) and revised labels (annotated by humans) in Figure 7(a). The revised labels were used to fine-tune the AI; we then compared the pseudo labels predicted by the AI before and after fine-tuning for the same CT volume from the JHH dataset, as shown in Figure 7(b). After fine-tuning the AI using the revised labels, the AI was able to accurately segment the entire stomach on the previously unseen CT volumes. This demonstrated the high quality of the revised labels and their efficacy in enhancing AI performance. We further quantified the human annotation quality measured by Dice Similarity Coefficient (DSC) and normalized surface Dice (NSD) using 1,000 CT volumes from the JHH dataset. The results in Appendix Table 5 show continual improvements in label quality along the active learning procedure. For example, AI models exhibited a marked improvement in the (pseudo) labels of aorta and postcava, jumping from 72.3% to 83.7% and 76.1% to 78.6%, respectively.

4.2 Label Bias and Segmentation Quality Evaluation

Evaluation of label bias. We first evaluate the predictions made by three AI architectures, i.e., Swin UNETR, nnU-Net, and U-Net, on CT volumes derived from the JHH dataset [92]. The segmentation predictions of these architectures are illustrated in Appendix Figure 10. We use the original annotations of the JHH dataset [92], which include our eight target organs. Consequently, for each CT volume, we have four predictions: three from the AI models and one from human experts. We then compute the DSC score between each pair across eight organs. These results are presented in Figure 8(a) and demonstrate that while the three AI architectures produce slightly varied predictions, they closely resemble those performed by human experts. To prevent label bias to a specific architecture, the final annotations in our AbdomenAtlas-8K are determined by averaging the three AI predictions. Our approach contrasts with the TotalSegmentator dataset [89], which exclusively relies on nnU-Net for segmentation. This could potentially generate a biased dataset that impedes model generalization across different architectures or scenarios.

Evaluation of AI prediction quality. To assess the automated annotation quality in AbdomenAtlas-8K, we asked for the assistance of two additional human annotators to help us modify the pseudo labels of our target eight organs. Due to time limitations, they are only able to revise six out of the eight organs, specifically on the 17 CT volumes from the BTCV dataset [45]. We compute DSC scores comparing AI predictions with those annotated independently by two human annotators, referred to as Dr1 and Dr2. Figure 8(b) presents the results for six organs. The DSC scores between AI predictions and each human annotator (Swin UNETR vs. Dr1 or Swin UNETR vs. Dr2) consistently exceed the score between the two human annotators (Dr1 vs. Dr2). These findings suggest that the automated AI annotations fall within the range of inter-annotator variability, thereby indicating that the quality of our automated annotations is comparable to human annotations.

4.3 Benchmarking

AbdomenAtlas-8K enables precision medicine for various downstream applications. We showcased one of the most pressing applications—early detection and localization of pancreatic cancer, an extremely deadly disease with a 5-year relative survival rate of only 12% in the United States. The AI trained on a large, private dataset at Johns Hopkins Hospital (JHH) performed arguably higher than typical radiologists [92, 46, 41]. But this AI model and annotated dataset were inaccessible due to the many policies. Now, our paper demonstrated that using AbdomenAtlas-8K (100% made up of publicly accessible CT volumes), AI can achieve similar performance when directly tested on the JHH dataset (see Table 2). This study is a concrete demonstration of how AbdomenAtlas-8K can be used to train AI models that can be generalized to many CT volumes from novel hospitals and be adapted to address a range of clinical problems. For a more in-depth analysis of the segmentation performance of AI models, we present the comprehensive category-wise scores comparison across eight organs in Appendix Table 7.

Table 2: **Benchmark results of AI models trained on AbdomenAtlas-8K.** We directly apply four AI models, i.e. SwinUNETR, UNETR, U-Net, and SegResNet, trained on AbdomenAtlas-8K (public) to JHH (unseen, private) and compared its performance with AI trained on JHH. The AI models trained on AbdomenAtlas-8K exhibit comparable results in average mDSC and mNSD across eight organs when compared with AI models trained on the JHH dataset, proving the high generalization capacity of AI models trained on AbdomenAtlas-8K.

AI Models	Trained on JHH (private)		Trained on AbdomenAtlas-8K (public)	
	mDSC (%)	mNSD (%)	mDSC (%)	mNSD (%)
SwinUNETR [84]	84.8 ± 12.6	66.5 ± 14.8	86.5 ± 7.5	60.8 ± 10.9
UNETR [27]	78.6 ± 13.0	53.9 ± 14.1	86.6 ± 6.5	59.5 ± 10.4
U-Net [73]	84.8 ± 11.2	64.2 ± 14.8	87.3 ± 6.0	61.0 ± 10.2
SegResNet [62]	87.6 ± 6.60	64.9 ± 10.9	87.0 ± 6.1	60.4 ± 10.1

5 Discussion

Impact. The scientific community has generally agreed that large volumes of annotated data are required for developing effective AI algorithms [38, 2, 59, 23, 26, 101]. For example, developing

Foundation Models for healthcare has recently raised much attention. The Foundation Model refers to an AI model that is trained on a large dataset and can be adapted to many specific downstream applications. This requires a large-scale, fully-annotated dataset. The currently available medical datasets are too small to represent the real data distribution in clinics [80, 12, 68]. The availability of large-scale, multi-center, fully-labeled data (summarized in Appendix Table 3) is one of the most significant cornerstones for both the development and evaluation of CADe systems [90, 103]. In recent years, the rise of imaging data archives [74, 56, 37, 3, 13, 15, 57, 97, 48, 99] and international competitions [44, 86, 78, 79, 29, 28, 1, 14, 4, 39] produced several publicly available datasets for benchmarking AI algorithms, but these datasets usually are of small size, contain partial labels, come from various scanners & protocols, and are therefore often limited in their scope [21, 49, 102, 95, 96, 77, 67, 88, 41]. We anticipate that our AbdomenAtlas-8K can play an important role in enabling the model to capture complex organ patterns, variations, and features across different imaging phases, modalities, and a wide range of populations. This has been partially evidenced in Appendix Table 6, wherein we generalize AI to CT volumes taken from different hospitals.

Limitation. Pseudo-labels have the potential to expedite tumor annotation procedures as well, but the risk of producing a large number of false positives is a significant concern. The presence of false positives could significantly increase the time annotators need to accept or reject a detection. In total, our AI models generate 51,852 tumor masks, including tumors in the colon, liver, hepatic vessel, pancreas, kidneys, and lung. However, only using AbdomenAtlas-8K, we are not able to assess the AI performance of tumor detection due to the lack of comprehensive pathology reports or expert annotations to describe the tumors in the majority of these publicly available CT volumes. To address this problem, we collected 392 tumor-free *(private)* CT volumes from Johns Hopkins Hospital to evaluate the false positives in the pseudo labels. Using kidney tumor detection as an example, there are 37 out of 392 CT volumes containing false positives (FPR = 9.4%) and a total of 161 false positives in the 37 CT volumes. Furthermore, we stratified the dataset based on the type of blood vessels, identifying average false positive rates of 11.83% and 4.6% for the venous and arterial vasculature, respectively. Therefore, while pseudo labels can be helpful, we anticipate a more effective method for false positive reduction is needed to enable a clinically practical tumor annotation, given the complexity of tumors compared with organs. As an extension, we plan to add tumor annotations to AbdomenAtlas-8K in three possible directions. **Firstly**, we plan to recruit more experienced radiologists to revise tumor annotations. **Secondly**, we will incorporate the pathology reports (based on biopsy results) into the human revision. These actions can reduce potential label biases and label errors from human annotators. **Thirdly**, we will exploit synthetic data (tumors) that can produce enormous tumor examples and their precise masks for AI training and validation [31, 46, 32].

In TotalSegmentator, the labels were largely generated by a single nnU-Net re-trained continually. Depending solely on nnU-Net could introduce a potential label bias favoring the nnU-Net architecture. This means that whenever TotalSegmentator is employed for benchmarking, nnU-Net would always outperform other segmentation architectures (e.g., UNETR, TransUNet, SwinUNETR, etc.). This observation has been made in several publications such as Huang et al. [34]. In contrast, our AbdomenAtlas-8K incorporates predictions from three different AI architectures, preventing bias towards one specific architecture [89]. However, such a solution comes with increased computational costs, and the performance of the AI architectures can vary. Taking an average of the predictions may result in the final outcome being pulled down by poorly performing AI architectures.

6 Conclusion

Our study shows the effectiveness of an active learning procedure, which combines the expertise of an annotator with the capabilities of a trained AI model. It not only deploys multiple AI models to detect prediction errors but also prompts annotators to revise these potential failures. Using this approach, we successfully annotated 8,448 abdominal CT volumes of different organs within three weeks, expediting the annotation process by a staggering factor of 533. Furthermore, our experiments demonstrate that the interaction between human and AI models can produce comparable or even superior results than that of a human annotator alone. This indicates that our approach can leverage the strengths of both human and AI models to achieve accurate and efficient annotation. Leveraging our efficient annotation framework, we anticipate that larger-scale medical datasets of various modalities, organs, and abnormalities can be curated at a significantly reduced cost, ultimately contributing to the development of Foundation Models in the medical domain [91, 51, 60].

Acknowledgments and Disclosure of Funding

References

[1] 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.

[2] Diego Ardila, Atilla P Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine*, 25(6):954–961, 2019.

[3] Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.

[4] Ujjwal Baid, Satyam Ghodasara, Michel Bilello, Suyash Mohan, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

[5] Maria-Florina Balcan, Andrei Broder, and Tong Zhang. Margin based active learning. In *International Conference on Computational Learning Theory*, pages 35–50. Springer, 2007.

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

[7] 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.

[8] Samuel Budd, Emma C Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, page 102062, 2021.

[9] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020.

[10] Jieneng Chen, Yingda Xia, Jiawen Yao, Ke Yan, Jianpeng Zhang, Le Lu, Fakai Wang, Bo Zhou, Mingyan Qiu, Qihang Yu, et al. Towards a single unified model for effective detection, segmentation, and diagnosis of eight major cancers using a large collection of ct scans. *arXiv preprint arXiv:2301.12291*, 2023.

[11] Liangyu Chen, Yutong Bai, Siyu Huang, Yongyi Lu, Bihan Wen, Alan L Yuille, and Zongwei Zhou. Making your first choice: To address cold start problem in vision active learning. In *Medical Imaging with Deep Learning*. 2023.

[12] Linda C Chu and Elliot K Fishman. Deep learning for pancreatic cancer detection: current challenges and future strategies. *The Lancet Digital Health*, 2(6):e271–e272, 2020.

[13] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26(6):1045–1057, 2013.

[14] Errol Colak, Felipe C Kitamura, Stephen B Hobbs, Carol C Wu, Matthew P Lungren, Luciano M Prevedello, Jayashree Kalpathy-Cramer, Robyn L Ball, George Shih, Anouk Stein, et al. The rsna pulmonary embolism ct dataset. *Radiology: Artificial Intelligence*, 3(2):e200254, 2021.

[15] Karen L Crawford, Scott C Neu, and Arthur W Toga. The image and data archive at the laboratory of neuro imaging. *Neuroimage*, 124:1080–1083, 2016.

[16] A Criminisi. Microsoft research cambridge (msrc) object recognition image database (version 2.0), 2004.

[17] Aron Culotta and Andrew McCallum. Reducing labeling effort for structured prediction tasks. In *AAAI*, volume 5, pages 746–751, 2005.

[18] Ido Dagan and Sean P Engelson. Committee-based sampling for training probabilistic classifiers. In *Machine Learning Proceedings 1995*, pages 150–157. Elsevier, 1995.

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[20] Yang Deng, Ce Wang, Yuan Hui, Qian Li, Jun Li, Shiwei Luo, Mengke Sun, Quan Quan, Shuxin Yang, You Hao, et al. Ctspine1k: A large-scale dataset for spinal vertebrae segmentation in computed tomography. *arXiv preprint arXiv:2105.14711*, 2021.

[21] Nanqing Dong, Michael Kampffmeyer, Xiaodan Liang, Min Xu, Irina Voiculescu, and Eric P Xing. Towards robust medical image segmentation on small-scale data with incomplete labels. *arXiv preprint arXiv:2011.14164*, 2020.

[22] Matthias Eisenmann, Annika Reinke, Vivienn Weru, Minu Dietlinde Tizabi, Fabian Isensee, Tim J Adler, Sharib Ali, Vincent Andrearczyk, Marc Aubreville, Ujjwal Baid, et al. Why is the winner the best? *arXiv preprint arXiv:2303.17719*, 2023.

[23] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

[24] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR, 2017.

[25] Eli Gibson, Francesco Giganti, Yipeng Hu, Ester Bonmati, Steve Bandula, Kurinchi Gurusamy, Brian Davidson, Stephen P Pereira, Matthew J Clarkson, and Dean C Barratt. Automatic multi-organ segmentation on abdominal ct with dense v-networks. *IEEE transactions on medical imaging*, 37(8):1822–1834, 2018.

[26] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

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

[28] Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical image analysis*, 67:101821, 2021.

[29] 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.

[30] Hideitsu Hino. Active learning: Problem settings and recent developments. *arXiv preprint arXiv:2012.04225*, 2020.

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

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

[33] Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu, Jiongquan Chen, Chaoyu Chen, et al. Segment anything model for medical images? *arXiv preprint arXiv:2304.14660*, 2023.

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

[35] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019.

[36] 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.

[37] Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008.

[38] Saurabh Jha and Eric J Topol. Adapting to artificial intelligence: radiologists and pathologists as information specialists. *Jama*, 316(22):2353–2354, 2016.

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

[40] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):1–8, 2019.

[41] Mintong Kang, Bowen Li, Zengle Zhu, Yongyi Lu, Elliot K Fishman, Alan L Yuille, and Zongwei Zhou. Label-assemble: Leveraging multiple datasets with partial labels. In *IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2023.

[42] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[43] Johannes Kulick, Robert Lieck, Marc Toussaint, et al. Active learning of hyperparameters: An expected cross entropy criterion for active model selection. *ArXiv e-prints*, 2014.

[44] Bennett Landman, Zhoubing Xu, Juan Eugenio Igelsias, Martin Styner, Thomas Robin Langerak, and Arno Klein. Multi-atlas labeling beyond the cranial vault-workshop and challenge. 2017.

[45] 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.

[46] 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.

[47] Xin Li and Yuhong Guo. Adaptive active learning for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 859–866, 2013.

[48] Yuexiang Li and Linlin Shen. Skin lesion analysis towards melanoma detection using deep learning network. *Sensors*, 18(2):556, 2018.

[49] Pengbo Liu, Li Xiao, and S Kevin Zhou. Incremental learning for multi-organ segmentation with partially labeled datasets. *arXiv preprint arXiv:2103.04526*, 2021.

[50] 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.

[51] Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023.

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

[53] Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Shihao Ma, Adamo Young, Cheng Zhu, Kangkang Meng, Xin Yang, Ziyan Huang, et al. Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge. *arXiv preprint arXiv:2308.05862*, 2023.

[54] 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.

[55] Dwarikanath Mahapatra, Behzad Bozorgtabar, Jean-Philippe Thiran, and Mauricio Reyes. Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 580–588. Springer, 2018.

[56] Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507, 2007.

[57] Mojtaba Masoudi, Hamid-Reza Pourreza, Mahdi Saadatmand-Tarzjan, Noushin Eftekhari, Fateme Shafiee Zargar, and Masoud Pezeshki Rad. A new dataset of computed-tomography angiography images for computer-aided detection of pulmonary embolism. *Scientific data*, 5(1):1–9, 2018.

[58] Andrew Kachites McCallumzy and Kamal Nigamy. Employing em and pool-based active learning for text classification. In *Proc. International Conference on Machine Learning*, pages 359–367. Citeseer, 1998.

[59] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788):89–94, 2020.

[60] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

[61] Prateek Munjal, Nasir Hayat, Munawar Hayat, Jamshid Sourati, and Shadab Khan. Towards robust and reproducible active learning using neural networks. *ArXiv, abs/2002.09564*, 2020.

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

[63] Vishwesh Nath, Dong Yang, Holger R Roth, and Daguang Xu. Warm start active learning with proxy labels and selection via semi-supervised fine-tuning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 297–308. Springer, 2022.

[64] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[65] S Park, LC Chu, EK Fishman, AL Yuille, B Vogelstein, KW Kinzler, KM Horton, RH Hruban, ES Zinreich, D Fadaei Fouladi, et al. Annotated normal ct data of the abdomen for deep learning: Challenges and strategies for implementation. *Diagnostic and interventional imaging*, 101(1):35–44, 2020.

[66] Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Gholamreza Reza Haffari, Anton Van Den Hengel, and Javen Qinfeng Shi. Active learning by feature mixing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12237–12246, 2022.

[67] Olivier Petit, Nicolas Thome, and Luc Soler. Iterative confidence relabeling with deep convnets for organ segmentation with partial labels. *Computerized Medical Imaging and Graphics*, page 101938, 2021.

[68] Deborah Plana, Dennis L Shung, Alyssa A Grimshaw, Anurag Saraf, Joseph JY Sung, and Benjamin H Kann. Randomized clinical trials of machine learning interventions in health care: a systematic review. *JAMA Network Open*, 5(9):e2233946–e2233946, 2022.

[69] Luciano M Prevedello, Safwan S Halabi, George Shih, Carol C Wu, Marc D Kohli, Falgun H Chokshi, Bradley J Erickson, Jayashree Kalpathy-Cramer, Katherine P Andriole, and Adam E Flanders. Challenges related to artificial intelligence research in medical imaging and the importance of image analysis competitions. *Radiology: Artificial Intelligence*, 1(1):e180031, 2019.

[70] Pranav Rajpurkar and Matthew P Lungren. The current and future state of ai interpretation of medical images. *New England Journal of Medicine*, 388(21):1981–1990, 2023.

[71] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *arXiv preprint arXiv:2009.00236*, 2020.

[72] 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.

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

[74] 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.

[75] Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active hidden markov models for information extraction. In *International Symposium on Intelligent Data Analysis*, pages 309–318. Springer, 2001.

[76] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

[77] Gonglei Shi, Li Xiao, Yang Chen, and S Kevin Zhou. Marginal loss and exclusion loss for partially supervised multi-organ segmentation. *Medical Image Analysis*, 70:101979, 2021.

[78] George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, 2019.

[79] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.

[80] Shelly Soffer, Avi Ben-Cohen, Orit Shimon, Michal Marianne Amitai, Hayit Greenspan, and Eyal Klang. Convolutional neural networks for radiologic images: a radiologist's guide. *Radiology*, 290(3):590–606, 2019.

[81] Jamshid Sourati, Ali Gholipour, Jennifer G Dy, Sila Kurugol, and Simon K Warfield. Active deep learning with fisher information for patch-wise semantic segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 83–91. Springer, 2018.

[82] Jamshid Sourati, Ali Gholipour, Jennifer G Dy, Xavier Tomas-Fernandez, Sila Kurugol, and Simon K Warfield. Intelligent labeling based on fisher information for medical image segmentation using deep learning. *IEEE transactions on medical imaging*, 38(11):2642–2653, 2019.

[83] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, page 101693, 2020.

[84] 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.

15

[85] Emily B Tsai, Scott Simpson, Matthew P Lungren, Michelle Hershman, Leonid Roshkovan, Errol Colak, Bradley J Erickson, George Shih, Anouk Stein, Jayashree Kalpathy-Cramer, et al. The rsna international covid-19 open radiology database (ricord). *Radiology*, 299(1):E204–E213, 2021.

[86] 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.

[87] Jingwen Wang, Yuguang Yan, Yubing Zhang, Guiping Cao, Ming Yang, and Michael K Ng. Deep reinforcement active learning for medical image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 33–42. Springer, 2020.

[88] Li Wang, Dong Li, Yousong Zhu, Lu Tian, and Yi Shan. Cross-dataset collaborative learning for semantic segmentation. *arXiv preprint arXiv:2103.11351*, 2021.

[89] Jakob Wasserthal, Manfred Meyer, Hanns-Christian Breit, Joshy Cyriac, Shan Yang, and Martin Segeroth. Totalsegmentator: robust segmentation of 104 anatomical structures in ct images. *arXiv preprint arXiv:2208.05868*, 2022.

[90] Martin J Willemink, Wojciech A Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R Folio, Ronald M Summers, Daniel L Rubin, and Matthew P Lungren. Preparing medical imaging data for machine learning. *Radiology*, 295(1):4–15, 2020.

[91] Martin J Willemink, Holger R Roth, and Veit Sandfort. Toward foundational deep learning models for medical imaging in the new era of transformer networks. *Radiology: Artificial Intelligence*, 4(6):e210284, 2022.

[92] 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.

[93] Yingda Xia, Qihang Yu, Linda Chu, Satomi Kawamoto, Seyoun Park, Fengze Liu, Jieneng Chen, Zhuotun Zhu, Bowen Li, Zongwei Zhou, Yongyi Lu, Yan Wang, Wei Shen, Lingxi Xie, Yuyin Zhou, Daniel Fouladi, Shahab Shayesteh, Scott Jefferson Graves, Alejandra Blanco, Eva Zinreich, Ken Kinzler, Ralph Gruban, Bert Vogelstein, Eliot Fishman, and Alan Yuille. Ai algorithms can assist radiologists in early detection of pancreatic neoplasms through venous and arterial ct imaging. In *Radiological Society of North America (RSNA)*, 2022.

[94] Binhui Xie, Longhui Yuan, Shuang Li, Chi Harold Liu, and Xinjing Cheng. Towards fewer annotations: Active learning via region impurity and prediction uncertainty for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8068–8078, 2022.

[95] Ke Yan, Jinzheng Cai, Adam P Harrison, Dakai Jin, Jing Xiao, and Le Lu. Universal lesion detection by learning from multiple heterogeneously labeled datasets. *arXiv preprint arXiv:2005.13753*, 2020.

[96] Ke Yan, Jinzheng Cai, Youjing Zheng, Adam P Harrison, Dakai Jin, You-bao Tang, Yu-Xing Tang, Lingyun Huang, Jing Xiao, and Le Lu. Learning from multiple datasets with heterogeneous and partial labels for universal lesion detection in ct. *IEEE Transactions on Medical Imaging*, 2020.

[97] Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of medical imaging*, 5(3):036501, 2018.

[98] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 399–407. Springer, 2017.

[99] Yang Yang, Xueyan Mei, Philip Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Katherine Link, Thomas Yang, Chendi Cao, Ying Wang, Hayit Greenspan, Timothy Deyer, and Zahi Fayad. Radimagenet: A large-scale radiologic dataset for enhancing deep learning transfer learning research, 2021.

[100] Yuan Yao, Fengze Liu, Zongwei Zhou, Yan Wang, Wei Shen, Alan Yuille, and Yongyi Lu. Unsupervised domain adaptation through shape modeling for medical image segmentation. *arXiv preprint arXiv:2207.02529*, 2022.

[101] Alan L Yuille and Chenxi Liu. Deep nets: What have they ever done for vision? *International Journal of Computer Vision*, 129(3):781–802, 2021.

[102] Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen. Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1195–1204, 2021.

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

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

[105] Zongwei Zhou, Jae Shin, Ruibin Feng, R Todd Hurst, Christopher B Kendall, and Jianming Liang. Integrating active learning and transfer learning for carotid intima-media thickness video interpretation. *Journal of digital imaging*, 32(2):290–299, 2019.

[106] Zongwei Zhou, Jae Shin, Lei Zhang, Suryakanth Gurudu, Michael Gotway, and Jianming Liang. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7340–7351, 2017.

[107] Zongwei Zhou, Jae Y Shin, Suryakanth R Gurudu, Michael B Gotway, and Jianming Liang. Active, continual fine tuning of convolutional neural networks for reducing annotation efforts. *Medical image analysis*, 71:101997, 2021.

Appendix

A   Implementation Details

Our AbdomenAtlas-8K includes three different AI architectures: Swin UNETR, U-Net, and nnU-Net. Swin UNETR and U-Net are trained on 2,100 CT volumes from 16 partially labeled public abdominal datasets (details in Figure 1). We use the Adam optimizer with an initial learning rate of $1e$-4, which is dynamically adjusted based on the combined dice and BCE losses. The weight decay is set to $1e$-5. nnU-Net is trained on 47 CT volumes from the BTCV [45] dataset. We utilize the SGD optimizer with an initial learning rate of $1e$-2 and a weight decay of $3e$-5. The AI architectures are trained on portal venous CT volumes. Eight NVIDIA RTX Quadro 8000 GPUs are used, and the batch size is 4 per GPU. The training process takes approximately 40 hours and the testing process takes around 12 hours.

B   Datasets and Permissions

We will only disseminate the annotations of the CT volumes separately, and users will retrieve the original CT volumes, if needed, from the original sources (websites). Everything we intend to create and license-out will be in separate files, and no modifications are necessary to the original CT volumes. The source and permissions are elaborated in Table 3. We have consulted with the lawyers at Johns Hopkins University, confirming the permissions of distributing the annotations based on the license of each dataset.

Dataset	Spleen	Liver	Kidney	Stomach
Multi-Atlas_Labeling	251.2 (211.0-401.7)	2331.4 (1743.9-2587.2)	466.3 (316.8-597.5)	431.9 (308.6-633.4)
TCIA_Pancreas-CT	224.3 (166.2-272.3)	1402.2 (1214.8-1687.3)	308.8 (276.9-372.3)	311.1 (228.3-398.5)
CHAOS	170.9 (118.4-234.2)	977.0 (896.0-1216.0)	212.0 (157.6-251.0)	130.1 (95.0-227.2)
LiTS	116.1 (69.6-193.1)	931.8 (573.5-1312.5)	206.9 (111.4-314.5)	142.7 (74.2-253.5)
KiTS	96.1 (54.6-202.0)	640.9 (421.4-1467.2)	186.8 (118.7-354.2)	107.3 (62.4-255.9)
WORD	178.1 (137.1-234.8)	1213.1 (1064.2-1406.9)	313.2 (267.1-361.0)	390.8 (251.5-518.9)
AbdomenCT-1K	54.2 (33.1-94.4)	364.3(255.8-689.2)	89.8 (61.6-154.7)	72.2 (41.8-133.2)
AMOS	268.6 (154.8-436.6)	1977.8 (1165.2-3127.8)	466.9 (276.4-723.1)	402.8 (226.2-748.2)
Decathlon	57.5 (35.7-105.4)	409.6 (262.8-671.7)	86.1 (55.2-140.6)	77.5 (48.5-137.6)
CT-ORG	245.0 (148.0-423.7)	2136.8 (1182.0-3108.4)	421.0 (230.2-727.6)	317.8 (151.2-565.3)
AbdomenCT-12organ	261.9 (199.9-383.5)	2111.0 (1890.2-2373.1)	523.6 (460.7-616.5)	414.8(276.3-561.0)
AbdomenAtlas-8K	79.6 (42.2-173.9)	544.4 (308.8-1216.0)	126.7 (70.8-285.9)	104.8 (54.2-232.4)

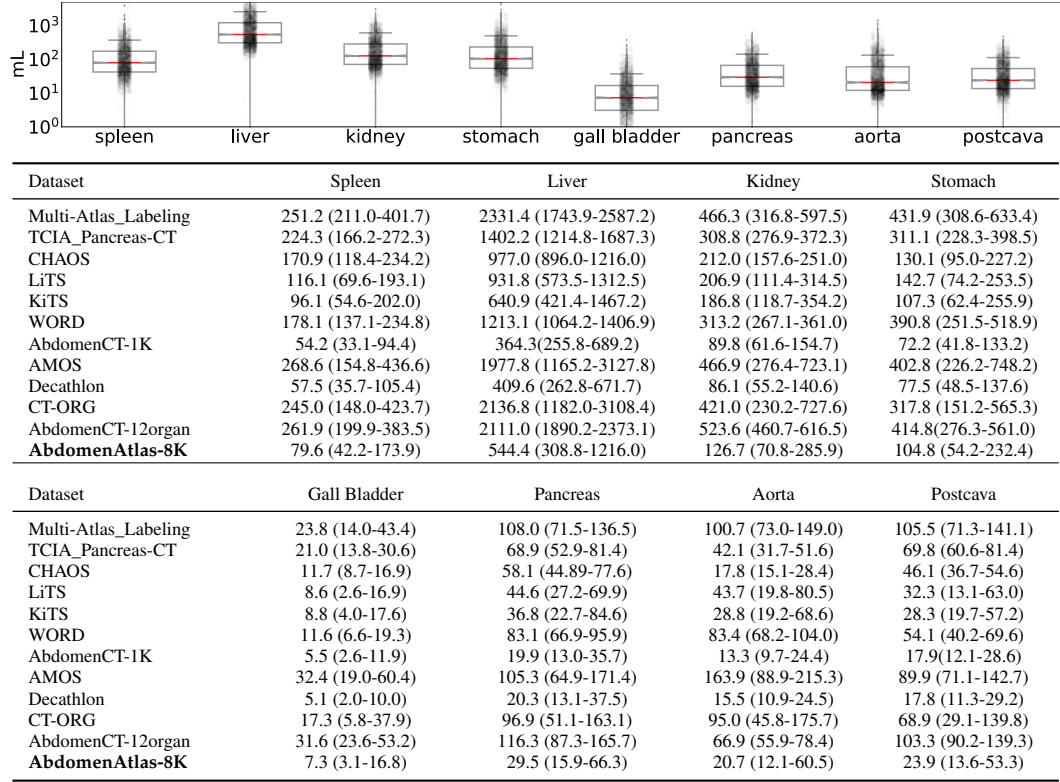| Dataset	Gall Bladder	Pancreas	Aorta	Postcava
Multi-Atlas_Labeling	23.8 (14.0-43.4)	108.0 (71.5-136.5)	100.7 (73.0-149.0)	105.5 (71.3-141.1)
TCIA_Pancreas-CT	21.0 (13.8-30.6)	68.9 (52.9-81.4)	42.1 (31.7-51.6)	69.8 (60.6-81.4)
CHAOS	11.7 (8.7-16.9)	58.1 (44.89-77.6)	17.8 (15.1-28.4)	46.1 (36.7-54.6)
LiTS	8.6 (2.6-16.9)	44.6 (27.2-69.9)	43.7 (19.8-80.5)	32.3 (13.1-63.0)
KiTS	8.8 (4.0-17.6)	36.8 (22.7-84.6)	28.8 (19.2-68.6)	28.3 (19.7-57.2)
WORD	11.6 (6.6-19.3)	83.1 (66.9-95.9)	83.4 (68.2-104.0)	54.1 (40.2-69.6)
AbdomenCT-1K	5.5 (2.6-11.9)	19.9 (13.0-35.7)	13.3 (9.7-24.4)	17.9(12.1-28.6)
AMOS	32.4 (19.0-60.4)	105.3 (64.9-171.4)	163.9 (88.9-215.3)	89.9 (71.1-142.7)
Decathlon	5.1 (2.0-10.0)	20.3 (13.1-37.5)	15.5 (10.9-24.5)	17.8 (11.3-29.2)
CT-ORG	17.3 (5.8-37.9)	96.9 (51.1-163.1)	95.0 (45.8-175.7)	68.9 (29.1-139.8)
AbdomenCT-12organ	31.6 (23.6-53.2)	116.3 (87.3-165.7)	66.9 (55.9-78.4)	103.3 (90.2-139.3)
AbdomenAtlas-8K	7.3 (3.1-16.8)	29.5 (15.9-66.3)	20.7 (12.1-60.5)	23.9 (13.6-53.3)

Figure 9: **The comparisons of volume distribution between public datasets and our dataset.** The values, expressed in mL, represent the median of the corresponding organ in the dataset, with the 25th and 75th percentiles indicated in parentheses.

Table 3: **The information for an assembly of datasets.** After summarizing the existing public datasets, we found that a combination of these datasets is fairly large (a total of 8,000 CT volumes and many more are coming over the years). The main challenge, however, is the absence of comprehensive annotations. In fact, scaling up annotations is much harder than scaling up CT volumes due to the limited time of expert radiologists. The existing labels of the public datasets are partial and incomplete, e.g., LiTS only had labels for the liver and liver tumors, and KiTS only had labels for the kidneys and kidney tumors. Our AbdomenAtlas-8K offered detailed annotations for all eight organs within each CT volume. With this, our contribution is significant.
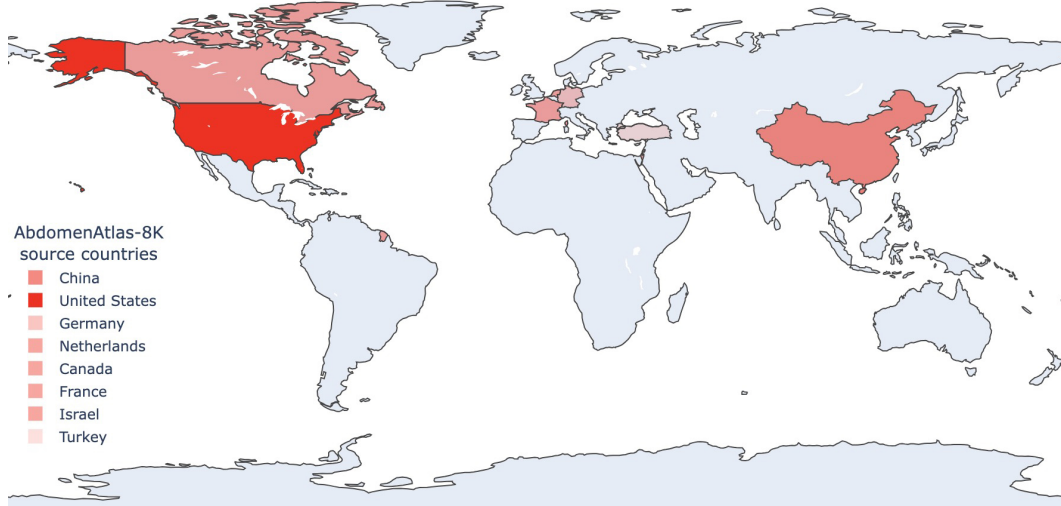
Dataset	# of organs	# of volumes	# of centers	source countries	license	# of volumes in AbdomenAtlas-8K
1. Pancreas-CT [74]	1	82	1	US	CC BY 3.0	42
2. LiTS [7]	1	201	7	DE, NL, CA, FR, IL	CC BY-SA 4.0	131
3. KiTS [29]	1	300	1	US	CC BY-NC-SA 4.0	300
4. AbdomenCT-1K [54]	4	1,112	12	DE, NL, CA, FR, IL, US, CN	CC BY-NC-SA	1000
5. CT-ORG [72]	5	140	8	DE, NL, CA, FR, IL, US	CC BY 3.0	140
6. CHAOS [86]	4	40	1	TR	CC BY-SA 4.0	20
7-11. MSD CT Tasks [1]	9	947	1	US	CC BY-SA 4.0	945
12. BTCV [45]	12	50	1	US	CC BY 4.0	47
13. AMOS22 [39]	15	500	2	CN	CC BY-NC-SA	200
14. WORD [50]	16	150	1	CN	GNU GPL 3.0	120
15. FLARE'23	13	4,000	30	-	CC BY-NC-ND 4.0	2200
16. AbdomenCT-12organ [53]	12	50	-	-	CC BY-NC-SA	50
17. AATTCT-IDS	-	300	1	CN	-	-
18. CTSpine1K	-	1,005	-	-	CC-BY-NC-SA	-
19. Abdominal Trauma Det.	-	3,147	23	-	-	-
20. TotalSegmentator [89]	104	1,204	1	CH	CC BY 4.0	-
21. JHH (*private*) [92]	20	5,282	1	US	JHU	-

US: United States   DE: Germany   NL: Netherlands   CA: Canada   FR: France   IL: Israel
CN: China   TR: Turkey   CH: Switzerland

Table 4: **The annotation time.** We employed two trained annotators to produce full annotations for the eight organs and recorded the average time needed for each organ. Both annotators have over 3-year experience in radiological image analysis.

Organs	Annotator #1	Annotator #2	Average
Spleen	30 min	28 min	29 min
Right Kidney	26 min	28 min	27 min
Left Kidney	23 min	25 min	24 min
Gall Bladder	11 min	10 min	10 min
Liver	40 min	40 min	40 min
Stomach	32 min	32 min	32 min
Aorta	22 min	20 min	21 min
Postcava (IVC)	27 min	30 min	28 min
Pancreas	19 min	19 min	19 min

Table 5: **Label quality improvement along the active learning procedure.** Assessing label quality is integral to ensure that the iterative training process is harnessing the most informative samples and that the revised labels are improved in quality. In doing so, we executed an external validation by leveraging a distinct dataset (JHH) separate from the active learning pool. Specifically, we trained AI models on the dataset at different stages: (1) before revising, (2) after revising Step #1, and (3) after revising Step #2. Subsequently, these three AI models were evaluated on the unseen JHH dataset. We randomly sampled 1,000 CT volumes from JHH and computed the mean DSC scores (mDSC) and normalized surface dice (mNSD) with a tolerance of 1mm for eight organs. After label revisions, both mDSC and mNSD showed an increase, demonstrating the effectiveness of the revised organ labels. Furthermore, we highlighted the two most significantly revised classes at each step in blue. For these organs, such as the aorta and postcava, AI models exhibited a marked improvement, jumping from 72.3% to 83.7% and 76.1% to 78.6%, respectively. Conversely, the segmentation performance for other classes saw only a slight uptick. This can be attributed to two key reasons: firstly, certain organs underwent minimal revisions at given steps, and secondly, some segmentation outcomes had already reached an apex of performance, exemplified by scores like DSC exceeding 90%.

Organ	Before revising		After revising Step #1		After revising Step #2	
	mDSC (%)	mNSD (%)	mDSC (%)	mNSD (%)	mDSC (%)	mNSD (%)
Spleen	94.0 ± 7.6	71.6 ± 10.5	94.1 ± 7.2	71.8 ± 11.5	94.7 ± 5.5	74.8 ± 9.5
Right Kidney	91.9 ± 8.0	64.8 ± 10.0	92.1 ± 7.0	65.0 ± 10.1	92.5 ± 7.0	64.7 ± 7.7
Left Kidney	90.8 ± 8.7	65.3 ± 10.1	91.0 ± 7.9	66.2 ± 10.1	91.5 ± 7.4	64.9 ± 9.8
Gall Bladder	82.4 ± 14.3	59.2 ± 17.2	79.8 ± 15.9	54.3 ± 17.6	81.7 ± 12.4	55.6 ± 15.1
Liver	95.5 ± 6.2	64.3 ± 9.5	95.7 ± 4.9	63.9 ± 9.2	95.8 ± 4.4	62.8 ± 8.2
Stomach	91.5 ± 8.0	52.7 ± 10.3	91.9 ± 6.8	53.9 ± 10.2	92.4 ± 5.6	53.0 ± 8.8
Aorta	72.3 ± 12.1	55.6 ± 11.3	82.9 ± 10.2	63.8 ± 11.5	83.7 ± 8.8	64.2 ± 9.8
Postcava (IVC)	76.1 ± 12.0	52.3 ± 11.0	77.2 ± 11.0	63.6 ± 10.7	78.6 ± 10.1	54.5 ± 9.5
Pancreas	79.4 ± 12.4	51.5 ± 10.1	80.1 ± 11.0	52.6 ± 9.9	79.2 ± 12.4	51.6 ± 9.3
Average	86.0 ± 9.9	59.7 ± 11.1	87.2 ± 9.1	60.6 ± 11.2	87.8 ± 8.2	60.7 ± 9.7

Table 6: **Assessing AI generalizability to multiple novel datasets.** On the left panel, the AI was trained on 16 partially labeled datasets; on the right panel, the AI was trained on the subset of AbdomenAtlas-8K. Note that the subset of AbdomenAtlas-8K has the same number of CT volumes (5,195), but their annotations are more comprehensive than 16 public datasets. We evaluated the segmentation performance of the two AI models on three unseen datasets with different sample sizes (N), i.e., FLARE'23 [52] ($N = 300$), TotalSegmentator [89] ($N = 1,204$) and JHH [92] ($N = 5,282$). The performance is measured by mean DSC scores (mDSC) and mean normalized surface dice (mNSD) with a tolerance of 1mm. The results showed that when pre-training the AI model using AbdomenAtlas-8K, the average mDSC and mNSD scores improved. This suggests that AbdomenAtlas-8K has higher quality compared to partially labeled abdominal datasets.

FLARE'23 [52]	Trained on 16 partially labeled datasets		Trained on AbdomenAtlas-8K	
	mDSC (%)	mNSD (%)	mDSC (%)	mNSD (%)
Spleen	96.1 ± 1.9	82.0 ± 16.4	$96.4 \pm 1.9^{****}$	$84.7 \pm 16.6^{****}$
Right Kidney	94.1 ± 6.7	81.5 ± 12.2	$96.3 \pm 1.9^{***}$	$80.0 \pm 14.1^{**}$
Left Kidney	93.9 ± 9.2	81.3 ± 12.7	93.1 ± 9.7^{ns}	$79.5 \pm 14.6^{****}$
Gall Bladder	85.1 ± 5.0	50.5 ± 16.3	$86.0 \pm 3.9^{***}$	$57.2 \pm 13.0^{****}$
Liver	97.3 ± 1.2	78.7 ± 15.8	$96.8 \pm 2.0^{****}$	$76.6 \pm 14.9^{***}$
Stomach	91.8 ± 5.0	49.2 ± 16.4	91.5 ± 6.3^{ns}	50.0 ± 17.3^{ns}
Aorta	83.7 ± 14.7	72.0 ± 18.4	81.8 ± 17.8^{ns}	73.4 ± 16.7^{ns}
Postcava (IVC)	85.7 ± 6.3	67.9 ± 8.8	86.3 ± 6.6^{ns}	$70.2 \pm 7.8^{**}$
Pancreas	85.3 ± 7.3	63.6 ± 15.2	85.5 ± 6.6^{ns}	$64.9 \pm 12.5^{*}$
Average	90.3 ± 6.4	69.6 ± 14.7	90.4 ± 8.2	70.7 ± 14.1

TotalSegmentator [89]	Trained on 16 partially labeled datasets		Trained on AbdomenAtlas-8K	
	mDSC (%)	mNSD (%)	mDSC (%)	mNSD (%)
Spleen	93.2 ± 11.0	59.4 ± 11.7	$94.8 \pm 8.4^{****}$	$65.0 \pm 12.3^{****}$
Right Kidney	91.2 ± 10.5	55.6 ± 13.9	$91.8 \pm 9.8^{**}$	$57.5 \pm 13.2^{***}$
Left Kidney	89.4 ± 14.2	58.7 ± 13.6	$90.4 \pm 12.7^{***}$	$59.9 \pm 13.3^{***}$
Gall Bladder	82.2 ± 15.5	45.2 ± 11.9	$82.9 \pm 15.9^{*}$	$48.3 \pm 12.6^{****}$
Liver	94.0 ± 10.0	53.1 ± 10.5	94.6 ± 9.3^{ns}	$54.4 \pm 10.8^{*}$
Stomach	82.8 ± 17.4	43.6 ± 11.9	$84.2 \pm 18.0^{**}$	$46.8 \pm 12.5^{****}$
Aorta	72.1 ± 19.9	47.9 ± 13.0	$75.7 \pm 19.5^{****}$	$53.5 \pm 14.2^{****}$
Postcava (IVC)	73.7 ± 18.9	47.5 ± 14.2	$74.7 \pm 18.7^{**}$	$50.2 \pm 15.3^{****}$
Pancreas	80.6 ± 15.4	40.9 ± 10.3	$82.4 \pm 15.7^{***}$	$46.5 \pm 10.8^{****}$
Average	84.4 ± 14.8	50.2 ± 12.3	85.7 ± 14.2	53.6 ± 12.8

JHH [92]	Trained on 16 partially labeled datasets		Trained on AbdomenAtlas-8K	
	mDSC (%)	mNSD (%)	mDSC (%)	mNSD (%)
Spleen	94.7 ± 1.8	72.0 ± 8.5	$95.0 \pm 1.5^{*}$	$74.4 \pm 8.4^{****}$
Right Kidney	92.2 ± 5.2	65.2 ± 9.3	92.2 ± 5.0^{ns}	$64.7 \pm 8.9^{**}$
Left Kidney	91.7 ± 1.8	65.7 ± 9.0	91.6 ± 1.8^{ns}	$65.1 \pm 9.3^{**}$
Gall Bladder	82.6 ± 12.6	60.8 ± 15.4	$83.8 \pm 10.5^{**}$	$60.1 \pm 15.8^{*}$
Liver	95.1 ± 7.7	66.0 ± 8.8	95.1 ± 7.7^{ns}	66.2 ± 8.6^{ns}
Stomach	91.6 ± 6.5	53.1 ± 9.8	$92.4 \pm 6.2^{**}$	$55.7 \pm 9.7^{****}$
Aorta	72.1 ± 10.2	55.5 ± 10.1	$74.2 \pm 10.5^{****}$	$57.5 \pm 10.3^{**}$
Postcava (IVC)	77.2 ± 11.0	54.9 ± 9.5	$77.7 \pm 10.9^{*}$	$57.5 \pm 10.3^{**}$
Pancreas	78.8 ± 12.4	52.3 ± 8.1	$79.3 \pm 11.3^{***}$	$55.2 \pm 8.8^{****}$
Average	86.2 ± 7.7	60.6 ± 9.8	86.8 ± 7.3	61.8 ± 9.9

^{ns} $P > 0.05$ * $P \leq 0.05$ ** $P \leq 0.01$ *** $P \leq 0.001$ **** $P \leq 0.0001$

Table 7: **Generalization of AI models trained on AbdomenAtlas-8K.** We directly applied AI trained on AbdomenAtlas-8K (public) to JHH (unseen, private) and compared its performance with AI trained on JHH.

SwinUNETR [84]	Trained on JHH (private)		Trained on AbdomenAtlas-8K (public)	
	mDSC (%)	mNSD (%)	mDSC (%)	mNSD (%)
Spleen	92.7 ± 12.2	76.7 ± 16.6	94.2 ± 4.7	73.1 ± 10.6
Right Kidney	92.6 ± 10.4	80.5 ± 11.6	90.8 ± 11.6	64.8 ± 10.9
Left Kidney	91.8 ± 9.2	78.8 ± 11.4	91.0 ± 4.3	65.3 ± 9.7
Gall Bladder	74.5 ± 23.9	54.1 ± 24.0	81.9 ± 13.4	57.2 ± 17.5
Liver	96.2 ± 1.1	69.5 ± 8.6	96.1 ± 1.0	65.2 ± 8.0
Stomach	91.7 ± 7.8	57.8 ± 14.4	93.1 ± 2.6	56.5 ± 10.1
Aorta†	87.6 ± 7.6	80.5 ± 11.7	73.1 ± 12.0	58.0 ± 12.7
Postcava (IVC)	70.4 ± 18.0	51.2 ± 17.2	77.9 ± 10.3	54.6 ± 10.5
Pancreas	65.8 ± 23.2	49.3 ± 17.4	81.1 ± 7.7	52.5 ± 8.1
Average	84.8 ± 12.6	66.5 ± 14.8	86.5 ± 7.5	60.8 ± 10.9

UNETR [27]	Trained on JHH (private)		Trained on AbdomenAtlas-8K (public)	
	mDSC (%)	mNSD (%)	mDSC (%)	mNSD (%)
Spleen	89.9 ± 8.2	62.3 ± 18.6	94.0 ± 4.6	72.2 ± 11.1
Right Kidney	79.3 ± 15.2	58.7 ± 16.4	92.0 ± 4.2	63.3 ± 8.7
Left Kidney	83.6 ± 13.4	63.2 ± 14.8	91.5 ± 2.2	64.2 ± 9.6
Gall Bladder	72.4 ± 17.2	50.3 ± 15.8	81.3 ± 12.1	55.9 ± 16.7
Liver	93.4 ± 3.7	55.7 ± 12.8	96.0 ± 0.9	64.2 ± 7.8
Stomach	84.9 ± 10.0	38.8 ± 9.8	92.9 ± 2.6	53.1 ± 9.5
Aorta†	83.9 ± 9.3	71.6 ± 14.1	74.0 ± 12.6	57.8 ± 12.8
Postcava (IVC)	58.3 ± 20.1	41.2 ± 13.0	76.2 ± 11.1	52.3 ± 10.1
Pancreas	62.1 ± 19.7	43.2 ± 11.6	81.2 ± 7.9	52.6 ± 7.5
Average	78.6 ± 13.0	53.9 ± 14.1	86.6 ± 6.5	59.5 ± 10.4

U-Net [73]	Trained on JHH (private)		Trained on AbdomenAtlas-8K (public)	
	mDSC (%)	mNSD (%)	mDSC (%)	mNSD (%)
Spleen	90.0 ± 12.5	72.0 ± 17.9	94.7 ± 2.8	74.0 ± 9.8
Right Kidney	92.3 ± 7.3	79.4 ± 11.2	91.8 ± 6.2	63.5 ± 9.2
Left Kidney	89.7 ± 10.4	74.4 ± 14.1	91.5 ± 2.6	65.7 ± 9.5
Gall Bladder	76.3 ± 17.7	56.0 ± 20.0	82.8 ± 11.2	58.5 ± 14.7
Liver	95.1 ± 2.4	63.7 ± 11.5	96.1 ± 0.9	64.2 ± 7.9
Stomach	90.8 ± 6.3	53.1 ± 13.4	93.2 ± 2.9	56.0 ± 9.8
Aorta†	84.6 ± 12.4	76.2 ± 14.7	74.3 ± 11.1	57.4 ± 12.4
Postcava (IVC)	73.0 ± 14.3	50.5 ± 17.4	78.2 ± 10.1	55.1 ± 10.5
Pancreas	71.2 ± 17.6	52.6 ± 12.8	82.7 ± 6.0	54.2 ± 8.0
Average	84.8 ± 11.2	64.2 ± 14.8	87.3 ± 6.0	61.0 ± 10.2

SegResNet [62]	Trained on JHH (private)		Trained on AbdomenAtlas-8K (public)	
	mDSC (%)	mNSD (%)	mDSC (%)	mNSD (%)
Spleen	95.2 ± 2.3	78.8 ± 9.5	94.6 ± 2.8	74.1 ± 9.8
Right Kidney	93.9 ± 6.5	80.3 ± 8.7	92.0 ± 5.9	65.5 ± 8.0
Left Kidney	92.7 ± 5.8	78.2 ± 8.7	91.8 ± 1.6	65.7 ± 7.9
Gall Bladder	80.4 ± 13.0	55.7 ± 16.4	82.8 ± 11.0	58.2 ± 16.3
Liver	90.2 ± 2.9	58.1 ± 7.5	95.8 ± 1.4	63.4 ± 8.3
Stomach	92.8 ± 5.1	55.2 ± 12.3	93.3 ± 2.9	55.6 ± 9.4
Aorta†	85.4 ± 6.9	75.2 ± 11.5	73.9 ± 12.2	55.5 ± 13.1
Postcava (IVC)	75.4 ± 9.8	45.5 ± 13.4	77.2 ± 10.6	53.2 ± 10.2
Pancreas	82.6 ± 6.8	56.9 ± 9.7	81.5 ± 6.5	52.6 ± 7.8
Average	87.6 ± 6.6	64.9 ± 10.9	87.0 ± 6.1	60.4 ± 10.1

† During the training process using AbdomenAtlas-8K, the ground truth of **Aorta** exhibits variation across different public datasets, resulting in the AI's inability to learn as anticipated. Conversely, when training on the JHH dataset, the annotation protocol of **Aorta** remains relatively consistent.
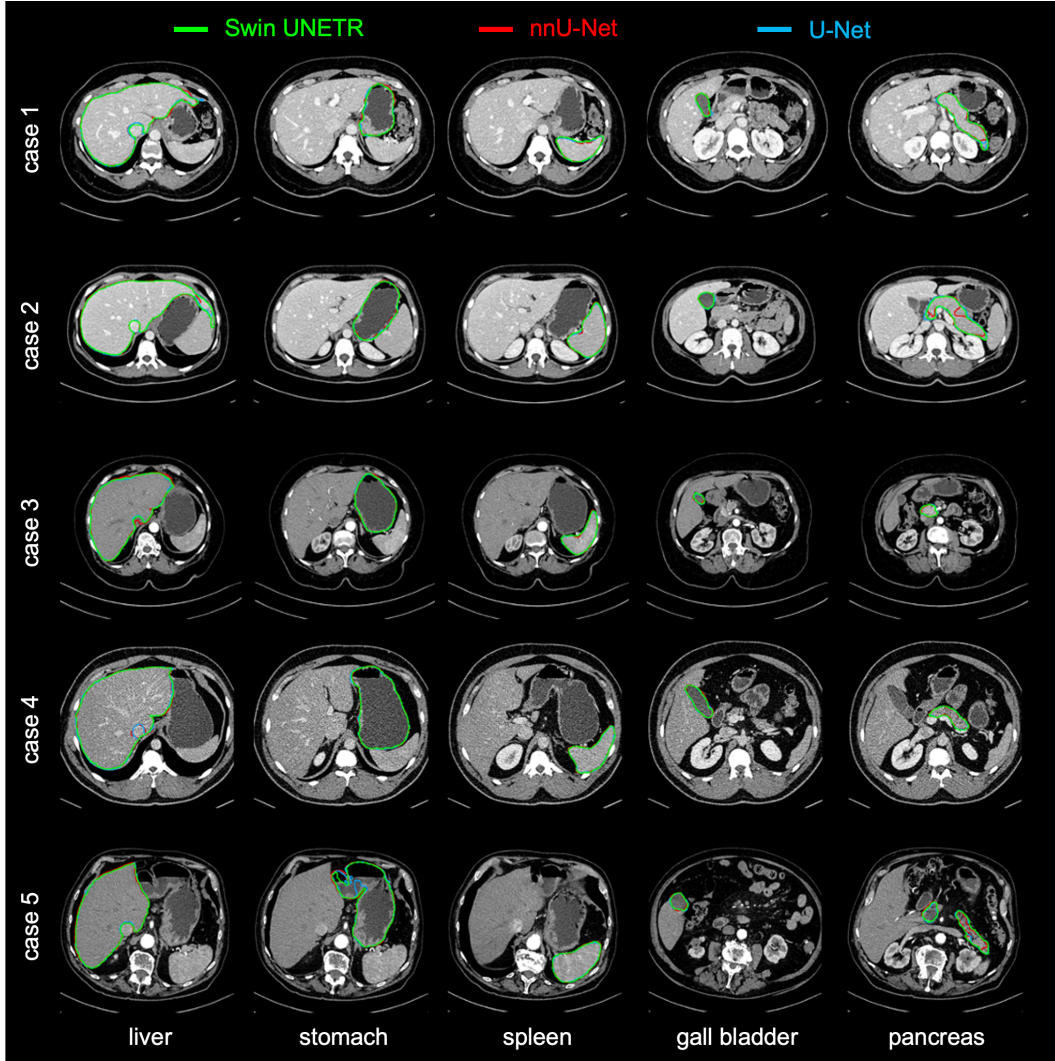
Figure 10: [Better viewed in color and zoomed in for details] **Variance in model predictions.** Our AbdomenAtlas-8K comprises an ensemble of three distinct segmentation architectures, thereby obviating the potential influence of architectural bias. As such, our AbdomenAtlas-8K stands out for its robustness and objectivity.
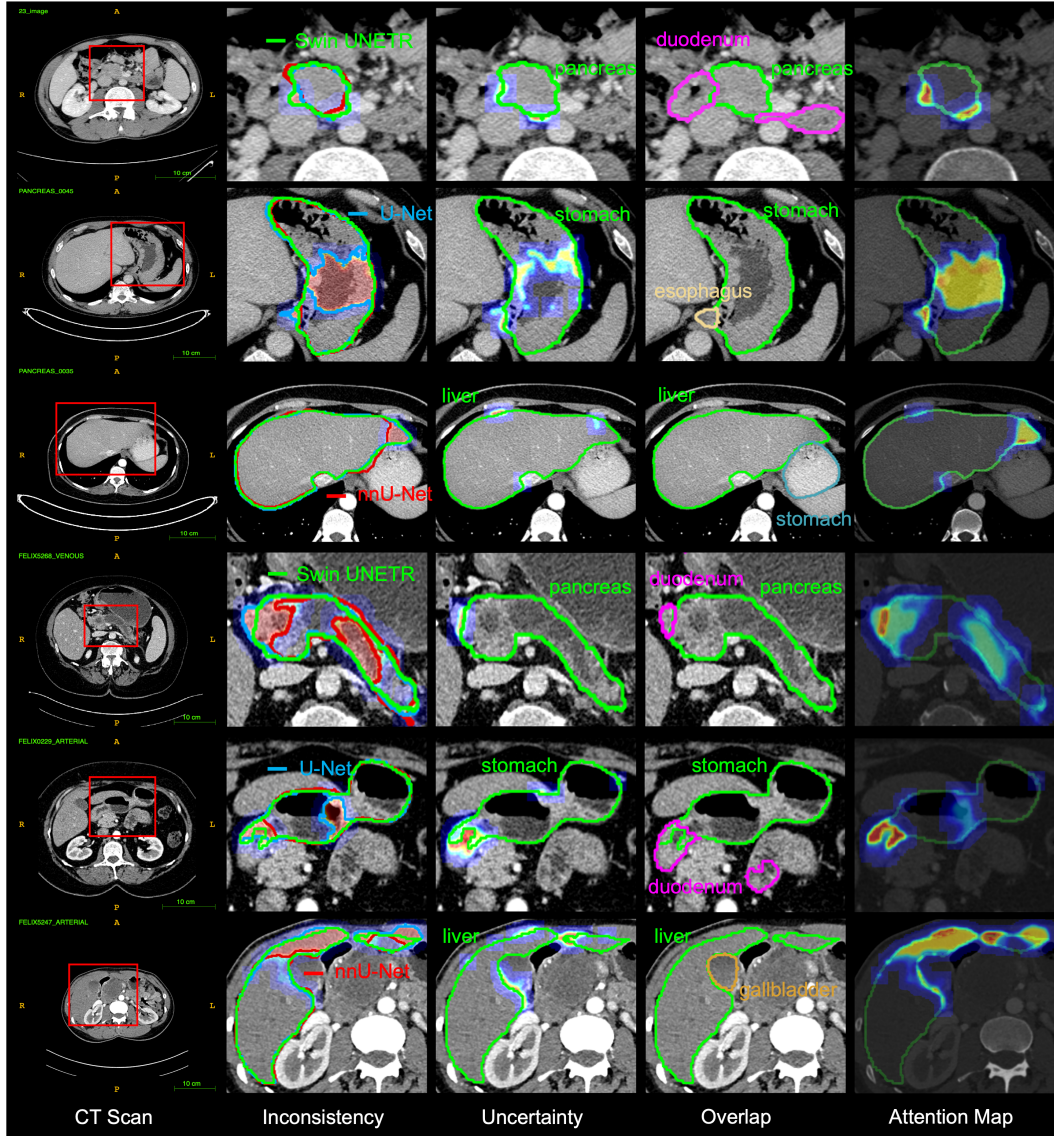
Figure 11: **Examples of attention map generation.** The attention map helps our annotator quickly and effectively locate regions with a high risk of prediction errors.

Figure 12: [Watch the animations by clicking them, better viewed with Acrobat/Foxit Reader] **Example annotated slices from AbdomenAtlas-8K**. AbdomenAtlas-8K includes annotations for eight organs, treating the left and right kidneys as a single organ.