EFFICIENT HUMAN-IN-THE-LOOP PANCREATIC TUMOR ANNOTATION VIA LARGE-SCALE PRE-TRAINED MODEL WITH ADAPTIVE POST-PROCESSING

Xinze Zhou[1,†], *Yuxuan Zhao*[2,†], *Chuntung Zhuang*[1], *Dexin Yu*[2], *Alan Yuille*[1], *Zongwei Zhou*[1,*]

[1]Johns Hopkins University [2]Qilu Hospital of Shandong University

Code & Data: https://github.com/ChrisXzz/EfficientAnno

ABSTRACT

Deep learning has significantly impacted fields like medical image analysis. However, the effectiveness of supervised learning models depends on large volumes of annotated data. Annotating medical images, especially for tumor regions, is costly and time-intensive. To overcome this limitation, this paper propose a novel framework to accelerate tumor annotation, producing a unique dataset with pancreatic lesions labeled in 300 CT scans from multiple centers. Traditional annotation methods would require an experienced radiologist around 216 hours for this task, whereas our framework completed it in 57 hours, achieving similar or even improved annotation quality. This success stems from two core elements: (1) pre-training the model on a large-scale multi-organ dataset, then fine-tuning it on the target domain before using the fine-tuned model to generate preliminary labels, and (2) adaptive post-processing based on statistical insights from an extensive in-house JHH dataset and clinical evidence. More importantly, our framework could help both the AI model and annotation processes, substantially reducing the annotation costs required to produce large-scale datasets for tumor detection and segmentation tasks for other organs, such as liver and kidney. With our framework, we introduce AbdomenAtlas2.0-Mini, a dataset of 300 CT scans with high-quality voxel-level pancreatic lesions annotations.

***Index Terms*—** Medical Image Analysis, Deep Learning, Human-in-the-loop Annotation

1. INTRODUCTION

Deep learning has achieved remarkable successes across diverse fields, including medical imaging, significantly advancing tasks such as organ segmentation [1] and tumor detection [2]. However, the performance of these models heavily relies on access to extensive labeled datasets. Data annotation in medical imaging, particularly for tasks such as tumor regions annotation, is a labor-intensive and time-consuming process [3]. Radiologists must manually label complex anatomical structures and tumor boundaries on volumetric images, a task that requires both expertise and precision. This process not only demands substantial time and resources but also limits the scalability of annotated datasets, which are crucial for training robust deep learning models. The high cost of annotation poses a significant barrier to developing large, diverse datasets, underscoring the need for efficient annotation frameworks that can reduce the burden on radiologists while maintaining high-quality annotations. Although current models trained on a single relatively small-scale dataset, such as MSD-Pancreas[4], perform well on the corresponding testing set, they often struggle to generalize across new or diverse datasets, particularly in the precise segmentation of tumor regions. This limitation, commonly known as the "generalization gap," [5] highlights a fundamental issue in deep learning applications for medical imaging. Addressing the generalization gap necessitates the use of larger datasets covering a wider range of protocols, thereby enabling models to learn representations that are robust across diverse domains. This strategy of leveraging large-scale, multi-protocol data aligns with the concept of scaling laws, which has recently been validated as an effective solution to improve model robustness and reduce the generalization gap in deep learning. [5]

To address these challenges, in this paper, we propose a new framework designed to streamline the annotation process for tumor lesions within CT scans. Our framework take advantage of large-scale supervised pre-trained model that has been validated to have excellent transfer learning capability in recent study [6] combined with adaptive post-processing, facilitating high-quality segmentation results with substantially reduced annotation times. Our empirical study demonstrated a 76% reduction in annotation time compared to manual annotation from scratch by radiologists. Radiologists only need to remove a small number of erroneous areas from the preliminary labels generated by the fine-tuned model after adaptive post-processing, with approximately 0.6 false positive regions per scan. This result suggests that our framework holds significant potential for scaling up annotation processes across broader datasets, enabling more efficient use of expert time. More importantly, our framework contributes to the creation of datasets that could foster improvements in tumor detection and segmentation models, facilitating more reliable applica-

† These authors contributed equally to this work.
* Correspondence to: Zongwei Zhou (zzhou82@jh.edu)

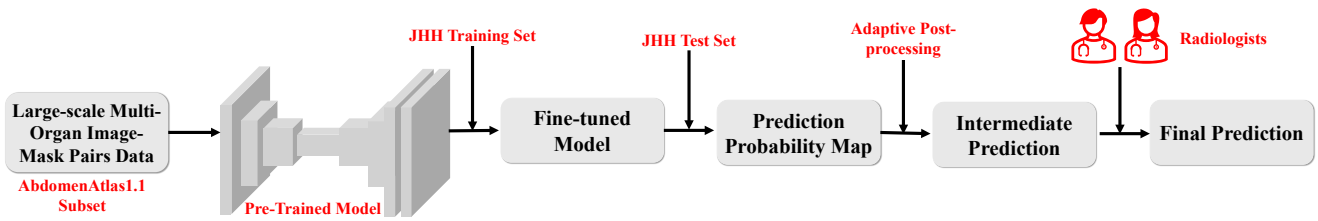

Fig. 1. Proposed framework: The process begins by pre-training a 3D TransU-Net model on a large-scale dataset, AbdomenAtlas 1.1. The pre-trained model is then fine-tuned on the JHH training set, a dataset of contrast-enhanced CT scans with pancreatic lesions. The fine-tuned model performs inference on the JHH test set, producing a probability map and determining a logit value corresponding to a sensitivity of 99%. Using this logit, an initial segmentation result is generated. Adaptive post-processing is subsequently applied to the initial segmentation, reducing false positives regions, to get intermediate prediction. Finally, radiologists review the post-processed prediction, revising it to produce the final prediction.

tions in clinical settings. Moreover, most existing AI-assisted annotation systems, such as MITK and 3D Slicer, rely on commercial software or require radiologists to learn how to use these tools from scratch, resulting in additional costs and unnecessary training time. In contrast, our proposed framework allows radiologists to directly refine AI-generated preliminary labels, avoiding these burdens. Based on our framework, we created and presented AbdomenAtlas2.0-Mini, a dataset of 300 CT scans with detailed per-voxel pancreatic tumors annotations, providing a good start and valuable resource for future studies.

2. MATERIALS AND METHODS

2.1. Supervised Pre-training and Fine-tuning

The overall workflow of our framework is illustrated in Fig. 1. Our framework utilizes a 3D TransU-Net [7] as the backbone model, selected for its effectiveness in volumetric segmentation tasks for medical imaging. This choice is flexible, as our framework can adapt to different backbone models or incorporate other advanced segmentation models—such as U-Net [8], Swin UNETR [9], or SegResNet [10]—depending on the specific task requirements and application scenarios. Our framework begins with a supervised pre-training phase on the subset of AbdomenAtlas1.1 dataset [6], a large, multi-organ dataset containing 2,100 fully annotated CT volumes that include detailed per-voxel annotations for 25 anatomical structures. This dataset serves as a comprehensive foundation, enabling the model to learn rich feature representations across diverse anatomical structures. To enhance the model's sensitivity to tumor characteristics, pseudo annotations for seven distinct tumor types were incorporated into the pre-training process. These annotations, generated through a mix of manual and semi-automated methods, provide varied representations of tumor morphology, further strengthening the model's feature extraction capabilities.

Following the pre-training phase, we fine-tuned the pre-trained 3D TransUNet model specifically for pancreatic lesions segmentation on a high-resolution CT dataset collected from Johns Hopkins Hospital (JHH) [2], which includes 5,176 CT volumes categorized into normal cases, as well as cases with pancreatic ductal adenocarcinoma (PDAC), cysts, and pancreatic neuroendocrine tumors (PNETs). For the fine-tuning phase, we utilized 3,159 CT scans from the training set, while the remaining 1,960 scans formed the test set, ensuring robust representation across all pancreatic lesions. After fine-tuning, the model, denoted as $f(\cdot)$, is capable of segmenting not only the pancreas itself but also the pancreatic duct and three types of pancreatic lesions, providing comprehensive predictions across these regions on new CT scans.

Methods	Pancreatic Lesions (PDAC, Cyst or PNET)			
	Sensitivity	Specificity	PPV	F1
QY-F Model [2]	92.4	90.5	95.2	93.8
SegResNet [10]	94.2	90.0	95.0	94.6
SuPreM [6]	94.2	92.7	96.3	95.3
3D TransUNet (Ours) [7]	95.2	95.7	98.0	96.6

Table 1. Comparison of pancreatic tumor detection results at the patient level, including QY-F Model, SegResNet, SuPreM, and our 3D TransUNet. Metrics reported include Sensitivity, Specificity, PPV, and F1-score, highlighting the superior performance of our 3D TransUNet.

Compared to the baseline QY-F model [2] and more current advanced models as shown in Table 1, our fine-tuned pre-trained model demonstrated superior detection performance for pancreatic tumors, achieving a sensitivity of 95.2% versus 92.4% for the QY-F model. We then construct the Free-response ROC (FROC) Curve using $f(\cdot, \theta)$ across a range of logit threshold values θ, which provide a trade-off map between sensitivity and false positive rate. Our experiments reveal a logit threshold **0.003**, achieving a tumor detection sensitivity of 99% while maintaining an acceptable false positive rate. To address the issue of false positives and enhance an-

notation efficiency, we implemented a series of adaptive post-processing methods to filter out most of the false positive regions.

2.2. Adaptive Post-processing

To further refine the model's preliminary predictions and enhance annotation efficiency, we applied a series of adaptive post-processing methods designed to filter out false positive regions and ensure that detected tumor regions closely align with anatomical characteristics of the pancreas. First of all, to ensure consistency in contrast across all CT scans we would handle with, we applied contrast normalization as a preliminary step, adjusting all CT intensities to a standardized range of (-1000, 1000) HU. Then, our post-processing pipeline includes four main steps, each tailored to address specific challenges in pancreatic lesions detection and segmentation:

- Removal of Non-Pancreatic Lesions: As an initial filtering step, we discarded any predicted tumor regions that lays outside the pancreas area.
- Tumor-to-Pancreas Volume Ratio Filtering: For each tumor category, we imposed a threshold based on the ratio of tumor volume to pancreas volume. Tumor regions exceeding this threshold would be reclassified as pancreas, which helps differentiate between actual tumors and tissue variations that may mimic tumors in appearance. These thresholds, derived from statistical analysis on the JHH dataset, are as follows: 133 for PDAC, 83 for cysts, and 46 for PNET. These values reflect the volume characteristics across a large cohort of pancreatic tumors cases, providing a reliable filter in our empirical study.
- Pancreatic Duct Intersection Filtering: To further refine the segmentation, any predicted cyst or PNET regions intersecting with the pancreatic duct would be reclassified as part of the pancreas. This step is essential for excluding regions that, while resembling tumors, are in fact pancreatic ductal extensions or irregularities, ensuring the specificity of tumor annotations.
- Contrast-Based Filtering: Under this standardized setting, we conducted statistical analyses on the JHH dataset, calculating the average HU values for regions within each tumor type. These thresholds, calculated from the JHH dataset, were set to capture the typical average HU values of the region for each tumor type: PDAC $(-206, 185)$, cysts $(-33, 222)$, and PNET $(-16, 270)$. Predicted regions falling outside these ranges were reclassified as pancreatic tissue. Given the large, representative sample size of our JHH dataset, we find these thresholds generalize well to most cases of pancreatic tumors. Our empirical study also validated this belief, such as successfully removing some fat regions that were incorrectly predicted as tumor areas

This comprehensive pipeline enables our framework to achieve high sensitivity in tumor detection while significantly reducing the annotation workload, making it an effective and scalable solution for data annotations.

2.3. Human-in-the-loop Annotation

To evaluate the effectiveness of our proposed framework in improving annotation efficiency, we firstly designed a comparative annotation study involving 100 CT scans collected from public datasets [1, 11, 12]. In this study, two senior radiologists with over 10 years of experience annotated the same CT images in two distinct settings: one radiologist annotated the images from scratch, while the other used our framework to assist with annotation. This comparison allowed us to quantify the time saved and annotation accuracy improvements achieved with our framework. Moreover, we further collected an additional 200 CT scans from public datasets, and annotated them using our framework to create the AbdomenAtlas2.0-Mini dataset.

3. RESULTS AND DISCUSSION

The results of our comparative annotation study underscore the effectiveness of the proposed framework in significantly enhancing annotation efficiency for radiologists. By employing our framework, the average annotation time per CT scan was reduced by almost 76%, demonstrating a substantial improvement in productivity. For instance, in cases with multiple lesions (two or more tumor regions), annotating from scratch required an average of 56 minutes, whereas using our framework reduced the annotation time to approximately 13 minutes. Based on our proposed framework, the original 216 hours required to create the AbdomenAtlas2.0-Mini dataset have been greatly reduced to 57 hours. In addition to the time reduction, the framework facilitates the consistent generation of detailed annotations, as illustrated in Figure 2, which shows several example cases from our newly created AbdomenAtlas2.0-Mini dataset. These examples highlight the quality and precision of the annotations achieved with our framework.

4. CONCLUSION

Enhancing annotation efficiency and accuracy in tumor regions presents a critical challenge in applying deep learning models to medical imaging. In this paper, we introduced a novel framework combining large-scale supervised pre-trained model with adaptive post-processing methods to streamline the annotation process. By leveraging this framework, we achieved a substantial reduction in annotation time for radiologists and created a high-quality annotated dataset, AbdomenAtlas2.0-Mini, setting the stage for creating even larger-scale datasets in the future. Our framework facilitates the efficient creation of high-quality datasets, which holds

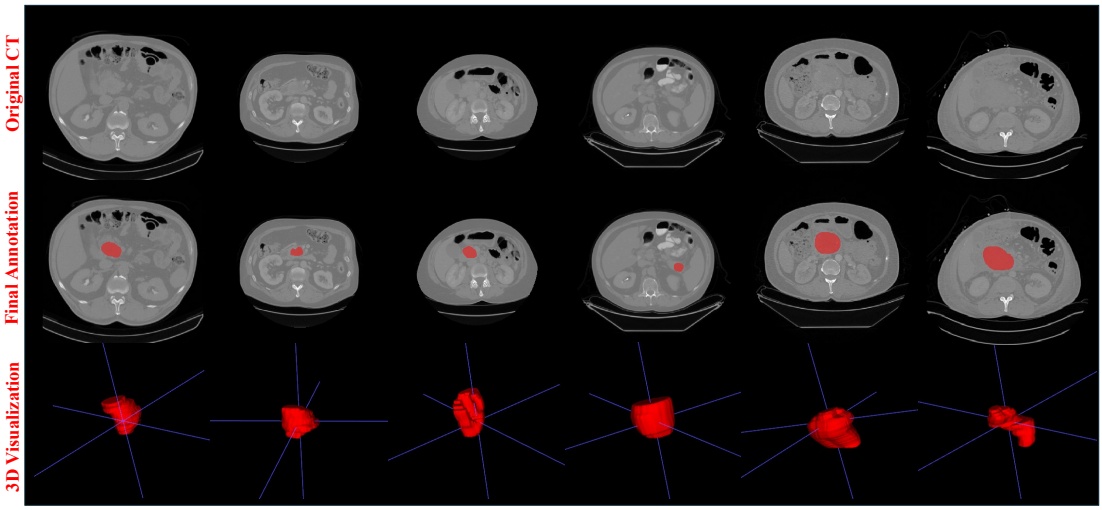

Fig. 2. **AbdomenAtlas2.0-mini annotations**. Red region denotes pancreatic lesions.

significant potential for improving the robustness and generalization of deep learning models, further supporting their practical application in clinical settings.

5. ACKNOWLEDGMENTS.

This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research and the McGovern Foundation.

6. REFERENCES

[1] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, Shucheng Cao, Qi Zhang, Shangqing Liu, Yunpeng Wang, Yuhui Li, Jian He, and Xiaoping Yang, "Abdomenct-1k: Is abdominal organ segmentation a solved problem?," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6695–6714, 2022.

[2] 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.

[3] Jun Ma, Yao Zhang, Song Gu, Xingle An, Zhihe Wang, Cheng Ge, Congcong Wang, Fan Zhang, Yu Wang, Yinan Xu, Shuiping Gou, Franz Thaler, Christian Payer, Darko Štern, Edward G.A. Henderson, Dónal M. McSweeney, Andrew Green, Price Jackson, Lachlan McIntosh, Quoc-Cuong Nguyen, Abdul Qayyum, Pierre-Henri Conze, Ziyan Huang, Ziqi Zhou, Deng-Ping Fan, Huan Xiong, Guoqiang Dong, Qiongjie Zhu, Jian He, and Xiaoping Yang, "Fast and low-gpu-memory abdomen ct organ segmentation: The flare challenge," *Medical Image Analysis*, vol. 82, pp. 102616, 2022.

[4] 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.

[5] Jee Seok Yoon, Kwanseok Oh, Yooseung Shin, Maciej A. Mazurowski, and Heung-Il Suk, "Domain generalization for medical image analysis: A survey," 2024.

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

[7] Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, Matthew P. Lungren, Shaoting Zhang, Lei Xing, Le Lu, Alan Yuille, and Yuyin Zhou, "Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers," *Medical Image Analysis*, vol. 97, pp. 103280, 2024.

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, Eds., Cham, 2015, pp. 234–241, Springer International Publishing.

[9] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R. Roth, and Daguang Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Alessandro Crimi and Spyridon Bakas, Eds., Cham, 2022, pp. 272–284, Springer International Publishing.

[10] Andriy Myronenko, "3d mri brain tumor segmentation using autoencoder regularization," 2018.

[11] Yuanfeng Ji, Haotian Bai, Jie Yang, Chongjian Ge, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, and Ping Luo, "Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation," 2022.

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