LEVERAGING AI PREDICTED AND EXPERT REVISED ANNOTATIONS IN INTERACTIVE SEGMENTATION: CONTINUAL TUNING OR FULL TRAINING?

Tiezheng Zhang [1] *Xiaoxi Chen*[2] *Chongyu Qu*[1] *Alan Yuille*[1] *Zongwei Zhou*[1,*]

[1]Johns Hopkins University [2]Shanghai Jiao Tong University

Code & Data: https://github.com/ollie-ztz/Continue_Tuning

ABSTRACT

Interactive segmentation, an integration of AI algorithms and human expertise, premises to improve the accuracy and efficiency of curating large-scale, detailed-annotated datasets in healthcare. Human experts revise the annotations predicted by AI, and in turn, AI improves its predictions by learning from these revised annotations. This interactive process continues to enhance the quality of annotations until no major revision is needed from experts. The key challenge is how to leverage *AI predicted* and *expert revised* annotations to iteratively improve the AI. Two problems arise: (1) The risk of catastrophic forgetting—the AI tends to forget the previously learned classes if it is only retrained using the expert revised classes. (2) Computational inefficiency when retraining the AI using both AI predicted and expert revised annotations; moreover, given the dominant AI predicted annotations in the dataset, the contribution of newly revised annotations—often account for a very small fraction—to the AI training remains marginal. This paper proposes *Continual Tuning* to address the problems from two perspectives: network design and data reuse. Firstly, we design a shared network for all classes followed by class-specific networks dedicated to individual classes. To mitigate forgetting, we freeze the shared network for previously learned classes and only update the class-specific network for revised classes. Secondly, we reuse a small fraction of data with previous annotations to avoid over-computing. The selection of such data relies on the importance estimate of each data. The importance score is computed by combining the uncertainty and consistency of AI predictions. Our experiments demonstrate that Continual Tuning achieves a speed 16× greater than repeatedly training AI from scratch without compromising the performance.

Index Terms— Interactive segmentation, Active learning

1. INTRODUCTION

Combining AI algorithms with human expertise in interactive segmentation [3, 4, 5] holds the promise of enhancing precision and productivity in the curation of large-scale, detailed annotated datasets such as SA-1B [6], TotalSegmentator [7], and AbdomenAtlas [8, 9]. During this synergy, human experts revise the AI predictions, and in return, AI enhances its predictions by adapting based on expert revised annotations. This iterative refinement continues until experts find that no substantial revisions are necessary [10, 11, 12, 13, 14].

However, the methodology to optimally leverage AI predicted annotations and expert revised annotations for the iterative enhancement of the AI remains ambiguous. There are two main issues to be considered. Firstly, there is the issue of catastrophic forgetting, which is shown in Figure 1 (a), where the AI often overlooks previously learned classes if it is exclusively retrained on expert revised annotations. Secondly, the process of retraining the AI using both its predictions and expert revised annotations is not only computationally demanding but also less impactful. This is because the AI predictions largely dominate the dataset, making the contribution of expert revised annotations—often a small portion—almost negligible in the training process. In addressing the phenomenon of catastrophic forgetting [15], one proposed strategy involves the retention of old class representations. For instance, Liu et al. [16] advocate for the preservation of prototypical representations across diverse classes. Similarly, Lao et al. [17] employ a feature replay methodology. Zhang et al. [18] use pseudo labels in their training process when the model is trained on new classes. However, these methods, which depend on the accuracy of annotations, might encounter practical challenges. For example, inconsistent or incomplete annotations can lead to the creation of misleading classes or the replay of incorrect features. Besides, Kirillov et al. [6] proposed to retrain the AI from scratch, a method we referred to as *Full Training*. However, this process could be time-consuming when applied to the medical domain. We seek to answer the following question: *Can we utilize the AI predicted and expert revised annotations effectively in interactive segmentation?*

To answer this question, we propose *Continual Tuning*, which focuses on two aspects: (i) network design and (ii) data reuse. **Firstly**, we develop a shared network that serves all classes, followed by different networks specifically designed for each class. To address the issue of forgetting, we keep the parameters of the shared network for the previously learned classes frozen while exclusively updating the network associated with the revised classes. As a result, the AI will not

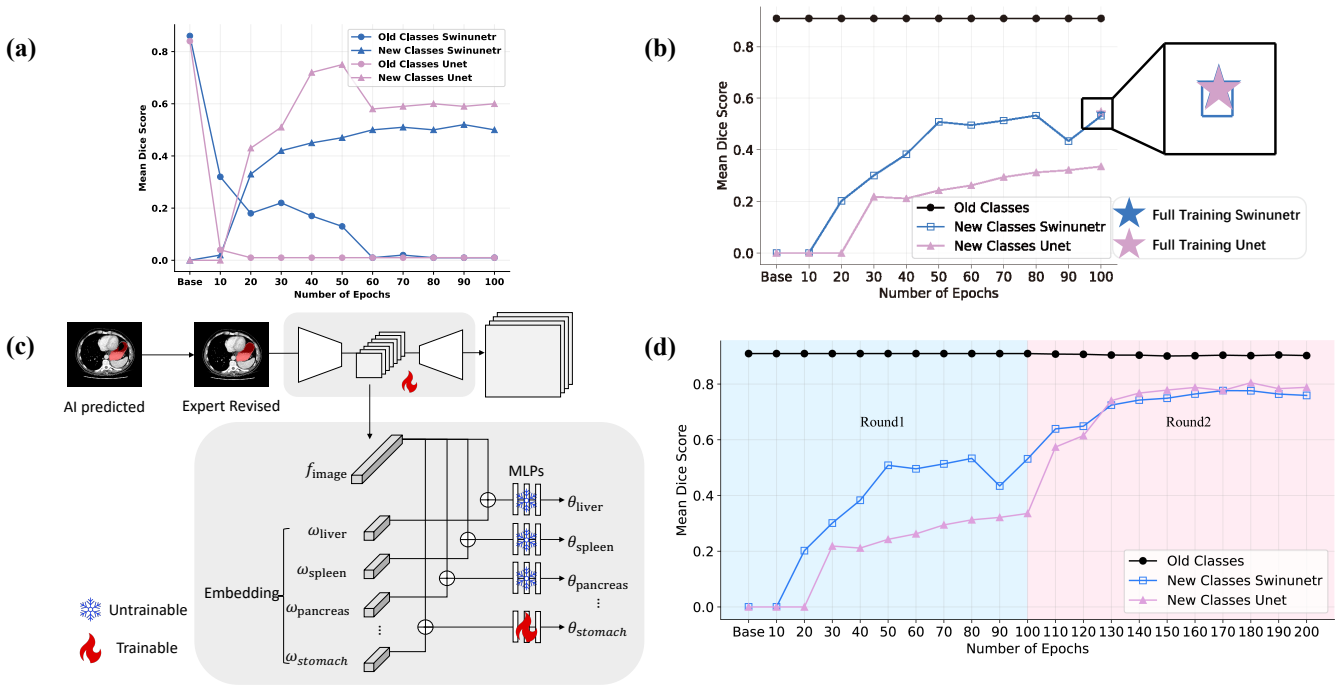

Fig. 1. **(a) Catastrophic gorgetting in Swin UNETR [1] and U-Net [2] backbones.** The old classes will be forgotten at the first few epochs when continual training AI models on data of new classes. **(b) Comparison of Continual Tuning and Full Training.** Two lines illustrate the mean DSC score using Continual Tuning method, while the asterisks show the final DSC score when applying Full Training. **(c) Shared Networks with Class-Specific Extensions.** The figure shows the networks we use, and we take the stomach as an example of the new class. **(d) Results of Continual Tuning on Two Rounds.** The blue region represents first-round results of Continual Tuning, and the red region, the second-round results.

forget the previously learned classes, as shown Figure 1 (b), while tuning only on the new classes revised by human experts. Additionally, Continual Tuning achieves a competitive Dice Similarity Coefficient (DSC) of 54.2% and 16 $\times$ faster than Full Training. **Secondly**, we reuse a small fraction of data with previous annotations to avoid over-computing. The selection of such data relies on the importance estimate [8, 19, 20] based on consistency, uncertainty, and overlapping. In summary, our ultimate goal is to continuously train AI models in interactive segmentation for better performance with the help of experts in the medical domain—this study makes a significant step towards it.

2. METHODOLOGY

Continual Tuning ideally enables efficient refinement of AI models using revised annotations. For instance, AI models should enhance their aorta segmentation performance when solely fine-tuned on revised aorta annotations. Thus, we have devised a shared network architecture that operates in conjunction with networks tailored for specific classes, as illustrated in Figure 1 (c). When fine-tuning AI models with expert revised annotations only, the shared network will remain unchanged, while the distinct networks associated with those revised annotations will be updated. With the help of text embeddings [21], which are encoded from the high-level visual semantics corresponding to each class, the class-specific networks become flexible to be updated. For instance, as depicted in Figure 1 (c), the AI models are fine-tuned exclusively with *stomach* annotations, and only the networks corresponding to the stomach are updated. In general, given the CT scans with revised annotations (X), the parameters of the corresponding MLP layer could be updated with :

$$\theta_k = MLP_k(E(X), \omega_k) \quad (1)$$

where $E(X)$ is the encoder feature of the image X, ω_k denotes the text embedding of each organ k. From the perspective of the data itself, given adequate computational resources, one can train AI models from scratch utilizing both AI predicted and expert-revised annotations, referred to as Full Training [6]. The improvement of the AI models could be slight due to the dominance of the unchanged annotations in the whole dataset. We propose to use expert-revised annotations in conjunction with AI predicted annotations (*Hybrid Data Continual Tuning*) to achieve significant improvements in AI models beyond just slight enhancements. Specifically, we express AI predicted annotations for each CT scan as

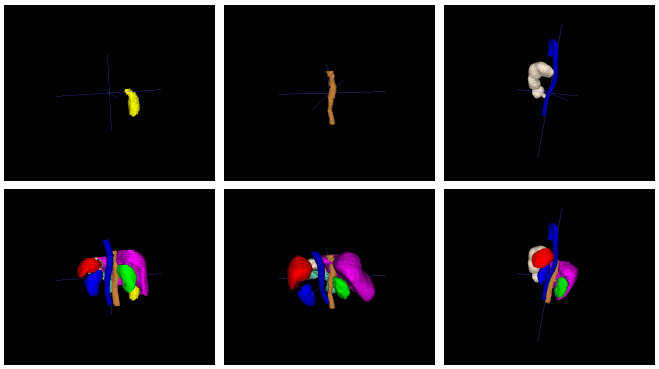

Fig. 2. **Examples of Hybrid Data.** In the upper row, the revised annotations for gall bladder, postcava (IVC), and stomach & aorta are presented from left to right. The lower row displays the corresponding hybrid annotations with old classes (liver, pancreas, left kidney, right kidney, and spleen).

$(C_1, C_2, C_3, ...C_n)$ where n is the total number of organs seen in this CT scan. The expert-revised annotations for each CT scan is $(C^*_1, C^*_2, C^*_3, ...C^*_m)$ where m is the number of organs revised by experts and $m \leq n$. By merging the revised annotations to the AI predicted annotations, the Hybrid Data could be expressed as$(C^*_1, C^*_2, C^*_3, ...C^*_m, C_n)$ shown in Figure 2. This design enables AI models to efficiently prioritize expert revised annotations without forgetting, due to the use of AI-predicted annotations for previous classes.

3. EXPERIMENTS & RESULTS & DISCUSSION

To prove that our class-specific model and Hybrid Data continual tuning can effectively work in interactive segmentation, we proposed three experiment settings: one is focused on the model trained from one dataset, the other one is using the model trained from 14 publicly available datasets, another one is the comparison between the previous two.

Implementation Details. The models were trained using the AdamW optimizer [22], coupled with a warm-up cosine scheduler lasting for 20 epochs [23], and a weight decay of $1e^{-5}$. For the learning rate (*lr*) and batch size, we opted for values of $1e^{-4}$ and 24, respectively. The pre-training phase extended over a total of 250 epochs. The training process was carried out across eight NVIDIA Quadro RTX 8000 cards.

3.1. Continual Tuning models pre-trained on one dataset

We used randomly selected 200 CT scans with annotations from the AbdomenCT-1K dataset [24] to train AI models with Swin UNETR [1] and U-Net [2] backbones. Those annotations comprise five classes: the liver, spleen, left kidney, right kidney, and pancreas. We asked an expert (over five years of experience) to annotate (using Pair) four classes: the stomach, postcava, aorta, and gall bladder in 12 out of the 200 CT scans, which we refer to as the first round of expert revised annotations. By contrast, we also used the same CT scans with nine classes to train the model from scratch, referred to as Full Training. After fine-tuning these revised annotations, the AI models are used to infer another 200 CT scans from the AbdomenCT-1K dataset. Then, 22 out of the 200 CT scans selected for revision in four classes (stomach, postcava, aorta, and gall bladder) are used for continual fine-tuning, referred to as the second round. The selection for the revision process is based on the uncertainty of the AI predicted annotations. To assess the performance of the models, we computed the DSC score on our proprietary JHH dataset containing high-quality annotations of all nine classes used in this experiment.

Results and Analysis. The quantitative results in Figure 1 (b) demonstrate that applying Continual Tuning on AI models could be 16× faster (200/12) compared with applying the Full Training method while still maintaining a similar DSC score (54.2% vs. 54.4%). The results in Figure 1 (d) further demonstrate the promise of Continual Tuning in interactive segmentation tasks. The first round in the blue region indicates that Continual Tuning assists in preventing the issue of forgetting. Then, the sharp increase from the first round part to the second round in red regions is attributed to the 22 CT scans predicted by the AI models after the first round of learning. There might be more prevalent errors in these 22 CT scans, which, when revised by the experts, can further enhance the model's performance. The final average DSC scores can achieve about 76.1% and 78.8% for Swin UNETR and U-Net backbones, respectively. We expect AI model performance to improve gradually through interactive segmentation and Continual Tuning.

3.2. Continual Tuning models pre-trained on 14 datasets

We first used 3,410 CT scans with annotations from 14 publicly available datasets to train the AI model with Swin UNETR backbones, which we refer to as the first round for this experiment. Those datasets are partially annotated but totally contain all nine classes used in the previous experiment. The AI model is used to infer another random 200 CT scans from the testing sets of 14 public datasets. Then, 12 out of the 200 CT scans selected for revision in four classes (stomach, postcava, aorta, and gall bladder) are used for continual fine-tuning, referred to as the second round. In this round, we tried to use three data strategies: one using revised annotations of 12 CT scans, the other one using 12 CT scans with nine classes, which is our Hybrid Data Continual Tuning method, and the last one is using all 200 CT scans with all nine classes.

Results and Analysis. One difference between this experiment and the previous is the scale of datasets used to train the model. We hypothesize that this kind of model is closer to the model used in real scenarios. From Table 1, we could find that if the model is only fine-tuned with the revised CT scans,

Organ	Before Fine-tuning mDice	Revised Data Only Continual Tuning mDice
Spleen	0.94	0.25
Right Kidney	0.92	0.08
Left Kidney	0.91	0.12
Pancreas	0.81	0.07
Liver	0.96	0.01
Stomach (11)	0.93	0.90
Aorta (12)	0.73	0.83
Postcava (IVC) (6)	0.76	0.75
Gall Bladder (1)	0.82	0.82
Organ	**Hybrid Data Continual Tuning mDice**	**Full Training mDice**
Spleen	0.95	0.94
Right Kidney	0.92	0.92
Left Kidney	0.91	0.91
Pancreas	0.82	0.82
Liver	0.96	0.96
Stomach (11)	0.93	0.93
Aorta (12)	0.83	0.75
Postcava(IVC) (6)	0.77	0.77
Gall Bladder (1)	0.82	0.82

Table 1. The numbers in parentheses indicate the amount of revised CT scans. The table illustrates the mean DSC score obtained from implementing various data strategies on the AI model that has been trained using 14 datasets.

Structures	Organ	Hybrid Data Continual Tuning mDice	Full Training mDice
Swin UNETR	Spleen	0.93	0.93
	Right Kidney	0.92	0.92
	Left Kidney	0.90	0.90
	Pancreas	0.73	0.80
	Liver	0.95	0.96
	Stomach (11)	0.77	0.89
	Aorta (12)	0.69	0.80
	Postcava(IVC) (6)	0.58	0.76
	Gall Bladder (1)	0.43	0.82
U-Net	Spleen	0.92	0.90
	Right Kidney	0.90	0.92
	Left Kidney	0.89	0.90
	Pancreas	0.79	0.81
	Liver	0.95	0.95
	Stomach (11)	0.64	0.86
	Aorta (12)	0.70	0.79
	Postcava(IVC) (6)	0.55	0.76
	Gall Bladder (1)	0.42	0.83

Table 2. The numbers in parentheses indicate the amount of revised CT scans. The table illustrates the mean DSC score obtained from implementing various data strategies on the AI models that have been trained using one dataset.

the model indeed improves the ability to segment the revised classes but also suffers from forgetting problems. Compared to using 200 CT scans in the second, using 12 CT scans could achieve a similar or better improvement of the model's ability. For example, the mean DSC score of the aorta improves by **10%** using Hybrid Data Continual Tuning, while it only improves by **2**% if we fine-tune the model with all 200 CT scans. This slight improvement is due to the dominance of unchanged data in the dataset (188 vs. 12).

3.3. Continual Tuning: Impact on Model Scales

We used 200 CT scans from one dataset with nine classes to train AI models with Swin UNETR and U-Net backbones. The AI models are used to infer another random 200 CT scans from the testing sets of this dataset. Then the same amount of CT scans are selected for revision. And we applied the same data strategies as we did in §3.2.

Results and Analysis. From Table 2, we could find that the models have better performance using all 200 CT scans, especially for organs that have revised annotations. Although the unchanged data still dominates the whole dataset, it does not weaken the influence of 12 revised annotations. The variations in phenotypes between §3.2 and §3.3 could be attributed to differences in dataset utilization. Multiple datasets could have different annotation principles. For example, some datasets include annotations for the stomach, including the cavity, while others may not. Although the annotation could be more accurate with the process of revision, the model's performance could be minimized by different annotation principles. On the other hand, if the models are trained on a single dataset, each organ follows a consistent annotation principle, and more data could lead to better performance.

4. CONCLUSION

In this paper, we propose Continual Tuning that integrates network design and data reuse to leverage AI predicted and expert revised annotations during the interactive segmentation procedure. Continual Tuning enables AI models to be fine-tuned efficiently (16$\times$ faster in our experiment) only with expert revised annotations in interactive segmentation tasks in the medical domain. This reveals the great potential for fine-tuning the AI models with incoming partial class datasets, e.g., AbdomenCT-1K, or datasets containing tumors.

Clinical Application. Our proposed Continual Tuning enhances diagnostic accuracy and minimizes annotation efforts, thus facilitating long-term learning and promoting trust in the model's decision-making process. This approach fosters continual improvement and the integration of the latest medical knowledge, thereby increasing the model's value in evidence-based healthcare settings.

Limitation. Continual Tuning involves several procedures that require human intervention, such as the annotation revision and selection process. This human involvement introduces a degree of subjectivity and variability, which may impact the overall quality and consistency of the annotations, consequently affecting the performance of the AI models. Secondly, the class-specific network we employ to prevent catastrophic forgetting is not inherently adaptive. As datasets evolve and new classes are introduced, the pre-defined class-specific network may become less effective.

Compliance with Ethical Standards. Committee/IRB of Johns Hopkins Medicine gave ethical approval for this work.

Acknowledgments. This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research and the McGovern Foundation. We thank Yaoyao Liu, Ju He for their constructive suggestions at several stages of the project.

5. REFERENCES

[1] 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*, 2022, pp. 20730–20740.

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[3] Sílvia D Olabarriaga and Arnold WM Smeulders, "Setting the mind for intelligent interactive segmentation: Overview, requirements, and framework," in *Biennial International Conference on Information Processing in Medical Imaging*. Springer, 1997, pp. 417–422.

[4] Sílvia Delgado Olabarriaga and Arnold WM Smeulders, "Interaction in the segmentation of medical images: A survey," *Medical image analysis*, vol. 5, no. 2, pp. 127–142, 2001.

[5] Feng Zhao and Xianghua Xie, "An overview of interactive medical image segmentation," *Annals of the BMVA*, vol. 2013, no. 7, pp. 1–22, 2013.

[6] 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.

[7] 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.

[8] Chongyu Qu, Tiezheng Zhang, Hualin Qiao, Jie Liu, Yucheng Tang, Alan Yuille, and Zongwei Zhou, "Abdomenatlas-8k: Annotating 8,000 abdominal ct volumes for multi-organ segmentation in three weeks," *Conference on Neural Information Processing Systems*, 2023.

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

[10] 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*, 2017, pp. 7340–7351.

[11] 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*, vol. 32, no. 2, pp. 290–299, 2019.

[12] 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*, vol. 71, pp. 101997, 2021.

[13] 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.

[14] Yaoyao Liu, Bernt Schiele, and Qianru Sun, "Adaptive aggregation networks for class-incremental learning," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2021, pp. 2544–2553.

[15] Stephan Lewandowsky and Shu-Chen Li, "Catastrophic interference in neural networks: Causes, solutions, and data," in *Interference and inhibition in cognition*, pp. 329–361. Elsevier, 1995.

[16] Pengbo Liu, Xia Wang, Mengsi Fan, Hongli Pan, Minmin Yin, Xiaohong Zhu, Dandan Du, Xiaoying Zhao, Li Xiao, Lian Ding, et al., "Learning incrementally to segment multiple organs in a ct image," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 714–724.

[17] Qicheng Lao, Xiang Jiang, Mohammad Havaei, and Yoshua Bengio, "A two-stream continual learning system with variational domain-agnostic feature replay," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 9, pp. 4466–4478, 2021.

[18] Yixiao Zhang, Xinyi Li, Huimiao Chen, Alan L Yuille, Yaoyao Liu, and Zongwei Zhou, "Continual learning for abdominal multi-organ and tumor segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2023, pp. 35–45.

[19] Xin Li and Yuhong Guo, "Adaptive active learning for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 859–866.

[20] Yarin Gal, Riashat Islam, and Zoubin Ghahramani, "Deep bayesian active learning with image data," in *International conference on machine learning*. PMLR, 2017, pp. 1183–1192.

[21] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou, "Clip-driven universal model for organ segmentation and tumor detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21152–21164.

[22] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[23] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He, "Accurate, large minibatch sgd: Training imagenet in 1 hour," *arXiv preprint arXiv:1706.02677*, 2017.

[24] 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*, vol. 44, no. 10, pp. 6695–6714, 2021.