Less Is More: A Comparison of Active Learning Strategies for 3D Medical Image Segmentation

Josafat-Mattias Burmeister[1*]   JOSAFAT-MATTIAS.BURMEISTER@STUDENT.HPI.DE
Marcel Fernandez Rosas[1*]   MARCEL.FERNANDEZROSAS@STUDENT.HPI.DE
Johannes Hagemann[1]   JOHANNES.HAGEMANN@STUDENT.HPI.DE
Jonas Kordt[1]   JONAS.KORDT@STUDENT.HPI.DE
Jasper Blum[1]   JASPER.BLUM@STUDENT.HPI.DE
Simon Shabo[1]   SIMON.SHABO@STUDENT.HPI.DE
Benjamin Bergner[1†]   BENJAMIN.BERGNER@HPI.DE
Christoph Lippert[1,2†]   CHRISTOPH.LIPPERT@HPI.DE

[1] *Digital Health & Machine Learning, Hasso Plattner Institute, University of Potsdam, Germany*
[2] *Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, NYC, USA*
* equal contribution, † equal advising

Abstract

Since labeling medical image data is a costly and labor-intensive process, active learning has gained much popularity in the medical image segmentation domain in recent years. A variety of active learning strategies have been proposed in the literature, but their effectiveness is highly dependent on the dataset and training scenario. To facilitate the comparison of existing strategies and provide a baseline for evaluating novel strategies, we evaluate the performance of several well-known active learning strategies on three datasets from the Medical Segmentation Decathlon. Additionally, we consider a strided sampling strategy specifically tailored to 3D image data. We demonstrate that both random and strided sampling act as strong baselines and discuss the advantages and disadvantages of the studied methods. To allow other researchers to compare their work to our results, we provide an open-source framework for benchmarking active learning strategies on a variety of medical segmentation datasets.

Keywords: Deep Learning, Active Learning, Semantic Segmentation, Medical Imaging

1. Introduction

Segmentation of organs and abnormalities in medical images is important for both research and clinical applications, such as radiotherapy planning or cardiac function assessment. In recent years, it has been shown that deep learning techniques can aid in these tasks, but they require a large amount of annotated training data (Lin et al., 2021; Chen et al., 2020). Especially in the medical domain, data annotation is very costly and time-consuming (da Cruz et al., 2021). One promising approach to reducing the labor-intensive annotation process is called active learning. In this approach, the goal is to only select the most informative samples for labeling. Previous work has shown that by training on an informative subset instead of the entire dataset, the number of required labeled samples can be reduced while achieving

similar performance (Zhang et al., 2019a,b). Various methods have been proposed to identify the most informative subset of a dataset. Common approaches are uncertainty sampling and representativeness sampling (Budd et al., 2021). Building on these approaches, several publications present strategies tailored to the domain of 3D medical image segmentation (Top et al., 2011; Konyushkova et al., 2015; Zhang et al., 2019a,b; Zheng et al., 2020). Commonly, these methods are evaluated by comparing them to random sampling, uncertainty sampling, or a model trained on the entire dataset. However, direct comparison of strategies is difficult because different authors use different datasets, model architectures, and active learning protocols. We seek to improve the comparability of active learning research in the 3D medical image segmentation domain through the following contributions:

1. We establish a baseline for evaluating active learning strategies by comparing the performance of several well-known query strategies and a strided sampling strategy on three 3D medical imaging datasets. We show that both random sampling and strided sampling act as strong baselines and discuss the advantages and disadvantages of the studied methods in 3D medical image segmentation.

2. We introduce an open-source benchmarking framework that allows to evaluate active learning strategies using a wide range of medical segmentation datasets.[1]

2. Related Work

In our work, we build upon previously proposed approaches to uncertainty sampling, representativeness sampling, as well as active learning with pseudo-labels.

Uncertainty sampling Approaches to capture uncertainty in a model's prediction include least confidence (Sharma et al., 2019), Shannon entropy (Konyushkova et al., 2015), and prediction variance (Wang et al., 2019; Zhang et al., 2019a). Our work evaluates the first two approaches, as they are among the best-known ones (Budd et al., 2021).

Representativeness sampling Common approaches to increase the diversity of a sampled dataset are to either select the data items that are most dissimilar to the already labeled items (Sharma et al., 2019) or that are most similar to the unlabeled items (Yang et al., 2017; Zheng et al., 2020). Image similarity is usually measured using feature vectors generated either by autoencoders (Zheng et al., 2020) or by the prediction models (Yang et al., 2017). We consider the latter approach because it is more efficient in terms of training cost.

Active learning with pseudo-labels Several authors propose to include the most confident predictions of a model as pseudo-labels in the training set (Gorriz et al., 2017; Zheng et al., 2020). To leverage the similarity of adjacent slices in medical 3D images, we propose a strided sampling approach in which pseudo-labels are generated through interpolation.

Benchmarking frameworks In the image classification domain, several extensible active learning frameworks exist, such as modAL (Danka and Horvath, 2018) or ALiPy (Tang et al., 2019). One main contribution of our work is the introduction of an active learning simulation framework for medical 3D image segmentation. Our framework is intended to be used as a benchmarking tool. It is readily extensible, experiments are comparable and reproducible.

---

1. *Active Segmentation framework:* `https://github.com/HealthML/active-segmentation`
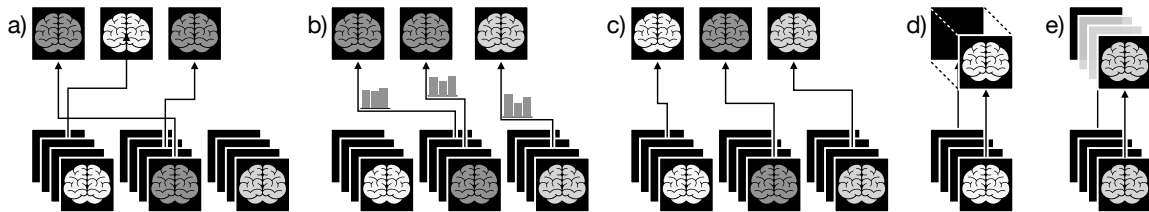
Figure 1: Visual representation of a) random sampling, b) uncertainty sampling, c) representativeness sampling, d) strided sampling, e) strided sampling with interpolation.

3. Methodology

3.1 Query Strategies

We evaluate two well-known types of active learning query strategies, uncertainty sampling and representativeness sampling. Additionally, we propose a strided sampling strategy, which exploits the three-dimensional structure of medical image data to generate pseudo-labels. Details on the query strategies are provided in Appendix A.

Uncertainty sampling In this approach, the current model is used to generate predictions on the set of unlabeled data. The data items where the predictions are most uncertain are selected for labeling. We consider two different methods to calculate the uncertainty, least confidence (LCUS) and Shannon entropy (EntrUS).

Representativeness sampling This approach aims to obtain a diverse and representative subset of a dataset. In our work, we consider two approaches that use model-generated feature vectors, namely distance-based (DistRS) and clustering-based (ClustRS) representativeness sampling. We also consider stratified random sampling (StrRS) as the most basic form of representativeness sampling, which ensures that the same number of 2D slices is sampled from each 3D scan.

Strided sampling & interpolation In this strategy, we consider sub-blocks of 3D scans, consisting of a top, a bottom, and intermediate 2D slices. Assuming that adjacent slices contain redundant information, only the top and the bottom slice of a block are selected for annotation. Intermediate slices are either not included in the training set, or pseudo-labels are generated by interpolating between the labels of the top and the bottom slice. Blocks are sampled either randomly or based on model uncertainty.

3.2 Datasets

We evaluate the query strategies on three tasks from the Medical Segmentation Decathlon challenge (Antonelli et al., 2021): Heart, hippocampus, and prostate segmentation. These three tasks were selected to cover segmentation tasks with varying complexity and datasets with disparate characteristics, i.e., different modalities, image and dataset sizes. Only the training datasets of these tasks are used in our study, as no labels are publicly available for the test datasets. To validate our results on unseen data, 20% of each training set was withheld for validation purposes.

3.3 Active Learning Protocol

We consider an active learning scenario in which 2D image slices are sampled from 3D image datasets and are annotated to train a 2D segmentation model. We prefer 2D over 3D models because initial experiments did not show a significant performance gap and the higher training speed of 2D models is better suited for interactive active learning scenarios. To train the initial models, 32 image slices are randomly sampled and annotated. Subsequently, 50 active learning iterations are performed. Each active learning iteration consists of a simulated labeling step and a training step. In the labeling step, 16 image slices are selected using a query strategy and annotated. In the training step, the model is trained for 10 epochs on the enlarged training dataset. To incrementally fit the model to the sampled data, we employ a fine-tuning approach in which model weights are not reset between active learning iterations. In all experiments, we use the 2D U-Net architecture (Ronneberger et al., 2015). Based on initial experiments, we use the Adam optimizer (Kingma and Ba, 2017) with a fixed learning rate of 10^{-4} and focal loss with $\gamma = 5$. Depending on image size, we choose batch sizes between 16 and 264. To ensure full reproducibility of our results, all random processes, such as model initialization or data shuffling, are seeded. Each experiment is repeated three times with different seeds. We choose the average and standard deviation of the mean validation Dice scores of these three runs as evaluation metrics. Details on our active learning framework are provided in Appendix B.

4. Results

Baseline strategies As shown in Fig. 2 and Fig. 3, most query strategies performed similarly on the three datasets studied, and no strategy outperformed random sampling (RandS) by a large margin. Which strategy worked best was highly dataset dependent: On the heart dataset, uncertainty sampling strategies slightly outperformed RandS in the early active learning iterations. DistRS slowed down the learning process significantly, while ClustRS produced a steeper learning curve than RandS. On the hippocampus dataset, both uncertainty sampling strategies resulted in flatter learning curves than RandS. All representativeness sampling strategies performed similarly. On the prostate dataset, EntrUS outperformed RandS in later iterations. StrRS also slightly outperformed RandS, whereas ClustRS sampling did not perform well.

Strided sampling & interpolation In this section, we consider random sampling-based strided sampling (StrideS) with and without interpolation. Results for further variants of strided sampling are provided in Appendix C. We consider block sizes 5, 10, and 15. StrideS without interpolation outperformed the RandS baseline for some combinations of block size and dataset, showing that it is already effective without generating pseudo-labels. On the hippocampus dataset, StrideS without interpolation exceeded RandS for all tested block sizes. On the heart dataset, StrideS outperformed RandS for block size 15, while on the prostate dataset, block size 5 produced the best results. Comparing StrideS with and without the generation of pseudo-labels showed that the inclusion of pseudo-labels barely improved model performance in most cases. For all three datasets, pseudo-label quality decreased with increasing block size. For the prostate dataset, e.g., the average interpolation Dice score was 0.88 ± 0.12 at block size 5, while it decreased to 0.81 ± 0.17 at block size 15.
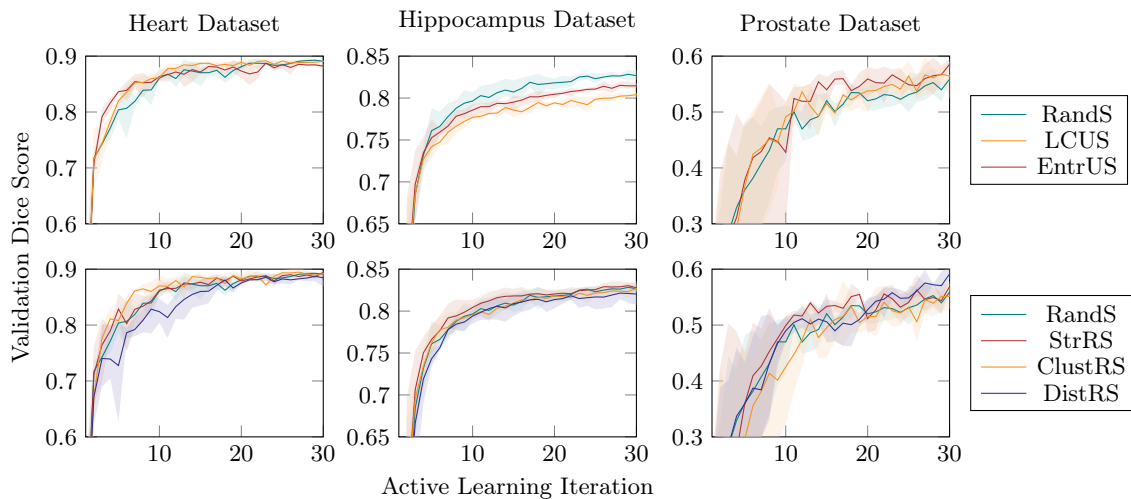
Figure 2: Performance of uncertainty sampling (first row) and representativeness sampling (second row) compared to a random sampling baseline.

Figure 3: Performance of strided sampling with block size 5 (first row), 10 (second row), and 15 (third row). The results shown were obtained with random block sampling and signed distance interpolation. For the hippocampus dataset, block size 15 was not tested since all scans in this dataset contain less than 15 slices.

5. Discussion

Our results show that the performance of uncertainty sampling and model-driven representativeness sampling strategies varies co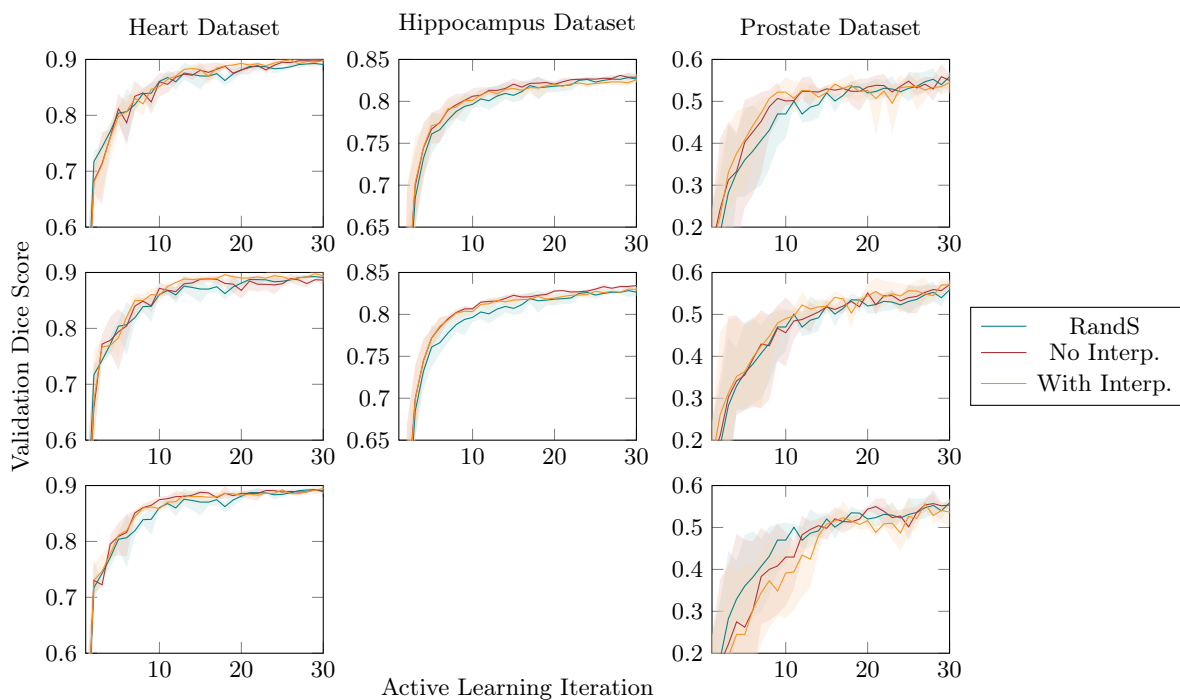nsiderably between datasets, while stratified random sampling and strided sampling yield decent results on all tested datasets. However, none of the tested strategies outperformed the random sampling baseline by a large margin. Our results are thus in line with the findings of Nath et al., who state that random sampling is a surprisingly difficult baseline to beat and that working active learning solutions are highly dataset dependent (Nath et al., 2021). In the following, we discuss reasons for this and consider the strengths of simple strategies such as stratified and strided sampling.

Uncertainty sampling The idea to select samples the model is most uncertain about is based on the idealized assumption that all samples contain independent, equally relevant information. When sampling 2D slices from 3D images, uncertainty sampling may result in selecting non-representative or redundant slices. To address this issue, a combination with representativeness sampling techniques might be beneficial. Another concern with uncertainty sampling is that neural networks tend to be overconfident, meaning that a model's predictions may not reflect its true uncertainty (Guo et al., 2017). Several approaches to this problem exist, such as model calibration (Wang et al., 2021), Bootstrapping (Yang et al., 2017), or the use of Bayesian models (Maddox et al., 2019), but not all have been studied in the context of 3D medical image segmentation.

Representativeness sampling Model-driven representativeness strategies rely on model-generated feature vectors. At an early stage of active learning, however, this feature representation might not be meaningful as the model is trained on very few samples initially. This issue could be addressed by pre-training models in a self-supervised manner. Model-independent strategies such as stratified random sampling can also be an alternative, as our results show.

Strided sampling & interpolation When appropriate block sizes were chosen, strided sampling outperformed random sampling on all datasets. Thus, it can act as a strong baseline for 3D medical image segmentation. Strided sampling has one strength in common with stratified random sampling: both strategies evenly sample 2D slices from the 3D scans ensuring that the entire dataset is covered. Additionally, strided sampling avoids the selection of neighboring, redundant slices. A trade-off must be considered when choosing the block size. Although the variance of the sampled data increases with larger blocks, resulting in a larger and more diverse dataset, the quality of the pseudo-labels decreases.

6. Conclusion

By comparing several basic active learning strategies, we have established a baseline for evaluating active learning strategies in the 3D medical image segmentation domain. Our results show that in the considered scenario, simple approaches such as random sampling or strided sampling are robust and may be sufficient for many use cases. Uncertainty sampling and model-driven representativeness sampling strategies seem to require more fine-tuning and possibly the integration of expert knowledge on dataset characteristics. By open-sourcing our evaluation framework, we aim to stimulate further research in this area.

References

Alexandra Albu, Trevor Beugeling, and Denis Laurendeau. A morphology-based approach for interslice interpolation of anatomical slices from volumetric images. *IEEE transactions on bio-medical engineering*, 55(8):2022–2038, 2008. ISSN 1558-2531. doi: 10.1109/TBME. 2008.921158. URL `https://pubmed.ncbi.nlm.nih.gov/18632365/`.

Mohamed Amgad, Habiba Elfandy, Hagar Hussein, Lamees A. Atteya, Mai A. T. Elsebaie, Lamia S. Abo Elnasr, Rokia A. Sakr, Hazem S. E. Salem, Ahmed F. Ismail, Anas M. Saad, Joumana Ahmed, Maha A. T. Elsebaie, Mustafijur Rahman, Inas A. Ruhban, Nada M. Elgazar, Yahya Alagha, Mohamed H. Osman, Ahmed M. Alhusseiny, Mariam M. Khalaf, Abo-Alela F. Younes, Ali Abdulkarim, Duaa M. Younes, Ahmed M. Gadallah, Ahmad M. Elkashash, Salma Y. Fala, Basma M. Zaki, Jonathan Beezley, Deepak R. Chittajallu, David Manthey, David A. Gutman, and Lee A. D. Cooper. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics*, 35(18):3461–3467, 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz083. URL `https://doi.org/10.1093/bioinformatics/btz083`.

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, AnnetteKopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Bram van Ginneken, Michel Bilello, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc J. Gollub, Stephan H. Heckers, Henkjan Huisman, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Jennifer S. Goli Pernicka, Kawal Rhode, Catalina Tobon-Gomez, Eugene Vorontsov, Henkjan Huisman, James A. Meakin, Sebastien Ourselin, Manuel Wiesenfarth, Pablo Arbelaez, Byeonguk Bae, Sihong Chen, Laura Daza, Jianjiang Feng, Baochun He, Fabian Isensee, Yuanfeng Ji, Fucang Jia, Namkug Kim, Ildoo Kim, Dorit Merhof, Akshay Pai, Beomhee Park, Mathias Perslev, Ramin Rezaiifar, Oliver Rippel, Ignacio Sarasua, Wei Shen, Jaemin Son, Christian Wachinger, Liansheng Wang, Yan Wang, Yingda Xia, Daguang Xu, Zhanwei Xu, Yefeng Zheng, Amber L. Simpson, Lena Maier-Hein, and M. Jorge Cardoso. The medical segmentation decathlon. *CoRR*, abs/2106.05735, 2021. doi: 10.48550/ARXIV.2106.05735. URL `https://arxiv.org/abs/2106.05735`.

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*, 71, 2021. ISSN 1361-8415. doi: 10.1016/j.media.2021.102062. URL `https://www.sciencedirect.com/science/article/abs/pii/S1361841521001080`.

Chen Chen, Chen Qin, Huaqi Qiu, Giacomo Tarroni, Jinming Duan, Wenjia Bai, and Daniel Rueckert. Deep learning for cardiac image segmentation: A review. *Frontiers in Cardiovascular Medicine*, 7, 2020. ISSN 2297-055X. doi: 10.3389/fcvm.2020.00025. URL `https://www.frontiersin.org/article/10.3389/fcvm.2020.00025`.

Leonardo C. da Cruz, César A. Sierra-Franco, Greis Francy M. Silva-Calpa, and Alberto Barbosa Raposo. Enabling autonomous medical image data annotation: A human-in-the-loop reinforcement learning approach. In Maria Ganzha, Leszek Maciaszek, Marcin Paprzycki, and Dominik Ślęzak, editors, *Proceedings of the 16th Conference on Computer Science and Intelligence Systems*, volume 25 of *FedCSIS 2021*,

pages 271–279. IEEE, 2021. ISBN 978-83-959183-6-0. doi: 10.15439/2021F86. URL `https://annals-csis.org/Volume_25/drp/pdf/86.pdf`.

Tivadar Danka and Peter Horvath. modAL: A modular active learning framework for Python. *CoRR*, abs/1805.00979, 2018. doi: 10.48550/arXiv.1805.00979. URL `https://arxiv.org/abs/1805.00979`.

William Falcon and The PyTorch Lightning team. PyTorch Lightning, 2019. URL `https://github.com/Lightning-AI/lightning`. Version 1.4.

Marc Gorriz, Axel Carlier, Emmanuel Faure, and Xavier Giró-i-Nieto. Cost-effective active learning for melanoma segmentation. *CoRR*, abs/1711.09168, 2017. doi: 10.48550/arXiv.1711.09168. URL `http://arxiv.org/abs/1711.09168`.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *ICML 2017*, page 1321–1330. JMLR.org, 2017. doi: 10.5555/3305381.3305518. URL `https://dl.acm.org/doi/10.5555/3305381.3305518`.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2017. doi: 10.48550/ARXIV.1412.6980. URL `https://arxiv.org/abs/1412.6980`.

Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. Introducing geometry in active learning for image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, ICCV 2015, pages 2974–2982. IEEE, 2015. ISBN 978-1-4673-8390-5. doi: 10.1109/ICCV.2015.340. URL `https://ieeexplore.ieee.org/document/7410697`.

Hui Lin, Haonan Xiao, Lei Dong, Kevin Boon-Keng Teo, Wei Zou, Jing Cai, and Taoran Li. Deep learning for automatic target volume segmentation in radiation therapy: a review. *Quantitative Imaging in Medicine and Surgery*, 11(12):4847–4858, 2021. ISSN 2223-4306. doi: 10.21037/qims-21-168. URL `https://qims.amegroups.com/article/view/81393`.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, ICCV 2017, pages 2999–3007. IEEE, 2017. ISBN 978-1-5386-1032-9. doi: 10.1109/ICCV.2017.324. URL `https://ieeexplore.ieee.org/document/8237586`.

Wesley J. Maddox, Timur Garipov, Pavel Izmailov, Dmitry Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. In Hanna Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32 of *NeurIPS 2019*. Curran Associates, 2019. ISBN 9781713807933. URL `https://proceedings.neurips.cc/paper/2019/file/118921efba23fc329e6560b27861f0c2-Paper.pdf`.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of the Fourth International Conference on 3D Vision*, 3DV 2016, pages 565–571. IEEE, 2016. ISBN

978-1-5090-5407-7. doi: 10.1109/3DV.2016.79. URL `https://ieeexplore.ieee.org/document/7785132`.

Vishwesh Nath, Dong Yang, Bennett A. Landman, Daguang Xu, and Holger R. Roth. Diminishing uncertainty within the training pool: Active learning for medical image segmentation. *IEEE Transactions on Medical Imaging*, 40(10):2534–2547, 2021. ISSN 1558-254X. doi: 10.1109/TMI.2020.3048055. URL `https://pubmed.ncbi.nlm.nih.gov/33373298/`.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In Hanna Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32 of *NeurIPS 2019*. Curran Associates, 2019. ISBN 9781713807933. URL `https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf`.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention*, MICCAI 2015, page 234–241. Springer, 2015. ISBN 978-3-319-24573-7. doi: 10.1007/978-3-319-24574-4_28. URL `https://link.springer.com/chapter/10.1007/978-3-319-24574-4_28`.

Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journa*, 27(3):379–423, 1948. ISSN 0005-8580. doi: 10.1002/j.1538-7305.1948.tb01338.x. URL `http://plan9.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf`.

Dhruv Sharma, Zahil Shanis, Chandan K. Reddy, Samuel Gerber, and Andinet Enquobahrie. Active learning technique for multimodal brain tumor segmentation using limited labeled images. In Qian Wang, Fausto Milletari, Hien V. Nguyen, Shadi Albarqouni, M. Jorge Cardoso, Nicola Rieke, Ziyue Xu, Konstantinos Kamnitsas, Vishal Patel, Badri Roysam, Steve Jiang, Kevin Zhou, Khoa Luu, and Ngan Le, editors, *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, MIL3ID 2019, page 148–156. Springer, 2019. ISBN 978-3-030-33391-1. doi: 10.1007/978-3-030-33391-1_17. URL `https://link.springer.com/chapter/10.1007/978-3-030-33391-1_17`.

Asim Smailagic, Pedro Costa, Hae Young Noh, Devesh Walawalkar, Kartik Khandelwal, Adrian Galdran, Mostafa Mirshekari, Jonathon Fagert, Susu Xu, Pei Zhang, and Aurélio Campilho. Medal: Accurate and robust deep active learning for medical image analysis. In M. Arif Wani, Mehmed Kantardzic, Moamar Sayed-Mouchaweh, Joao Gama, and Edwin Lughofer, editors, *Proceeding of the 17th IEEE International Conference on Machine Learning and Applications*, ICMLA 2018, pages 481–488. IEEE, 2018. ISBN 978-1-5386-6804-7. doi: 10.1109/ICMLA.2018.00078. URL `https://ieeexplore.ieee.org/document/8614103`.

Ying-Peng Tang, Guo-Xiang Li, and Sheng-Jun Huang. ALiPy: Active learning in python. *CoRR*, abs/1901.03802, 2019. doi: 10.48550/arXiv.1901.03802. URL `https://arxiv.org/abs/1901.03802`.

Andrew Top, Ghassan Hamarneh, and Rafeef Abugharbieh. Active learning for interactive 3d image segmentation. In Gabor Fichtinger, Anne Martel, and Terry Peters, editors, *Proceedings of the 14th International Conference on Medical Image Computing and Computer-Assisted Intervention*, MICCAI 2011, pages 603–610. Springer, 2011. ISBN 978-3-642-23626-6. doi: 10.1007/978-3-642-23626-6_74. URL `https://link.springer.com/chapter/10.1007/978-3-642-23626-6_74`.

Deng-Bao Wang, Lei Feng, and Min-Ling Zhang. Rethinking calibration of deep neural networks: Do not be afraid of overconfidence. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jenn Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34 of *NeurIPS 2021*, pages 11809–11820. Curran Associates, 2021. ISBN 9781713845393. URL `https://proceedings.neurips.cc/paper/2021/file/61f3a6dbc9120ea78ef75544826c814e-Paper.pdf`.

Jiacheng Wang, Zhaocai Chen, Liansheng Wang, and Qichao Zhou. An active learning with two-step query for medical image segmentation. In *Proceedings of the 2019 International Conference on Medical Imaging Physics and Engineering*, ICMIPE 2019. IEEE, 2019. ISBN 978-1-7281-4855-7. doi: 10.1109/ICMIPE47306.2019.9098222. URL `https://ieeexplore.ieee.org/document/9098222`.

Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z. Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. *CoRR*, abs/1706.04737, 2017. doi: 10.48550/arXiv.1706.04737. URL `http://arxiv.org/abs/1706.04737`.

Zhenxi Zhang, Jie Li, Zhusi Zhong, Zhicheng Jiao, and Xinbo Gao. A multi-resolution coarse-to-fine segmentation framework with active learning in 3d brain mri. In Zhen Cui, Jinshan Pan, Shanshan Zhang, Liang Xiao, and Jian Yang, editors, *Intelligence Science and Big Data Engineering. Visual Data Engineering*, IScIDE 2019, pages 285–298. Springer, 2019a. ISBN 978-3-030-36189-1. doi: 10.1007/978-3-030-36189-1_24. URL `https://link.springer.com/chapter/10.1007/978-3-030-36189-1_24`.

Zhenxi Zhang, Jie Li, Zhusi Zhong, Zhicheng Jiao, and Xinbo Gao. A sparse annotation strategy based on attention-guided active learning for 3d medical image segmentation. *CoRR*, abs/1906.07367, 2019b. doi: 10.48550/arXiv.1906.07367. URL `https://arxiv.org/pdf/1906.07367.pdf`.

Hao Zheng, Yizhe Zhang, Lin Yang, Chaoli Wang, and Danny Z. Chen. An annotation sparsification strategy for 3d medical image segmentation via representative selection and self-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34 of *AAAI 2020*, pages 6925–6932. Association for the Advancement of Artificial Intelligence, AAAI Press, 2020. ISBN 978-1-57735-835-0. doi: 10.1609/aaai.v34i04.6175. URL `https://ojs.aaai.org/index.php/AAAI/article/view/6175`.

Dženan Zukić, Jared Vicory, Matthew McCormick, Laura Wisse, Guido Gerig, Paul Yushkevich, and Stephen Aylward. Nd morphological contour interpolation. *The Insight Journal*, 2016. ISSN 2327-770X. doi: 10.54294/achtrg. URL `https://insight-journal.org/browse/publication/977`.

Appendix A.

In this supplementary material, implementation details of all query strategies evaluated in this work are provided.

Uncertainty Sampling

In uncertainty sampling, the data items where a model's predictions are most uncertain are selected for labeling. We consider two well-known methods for capturing model uncertainty: least confidence (Sharma et al., 2019) and Shannon entropy (Shannon, 1948; Top et al., 2011; Konyushkova et al., 2015).

Least confidence-based sampling In least confidence-based sampling, uncertainty is computed as follows: For each slice $x \in \mathbb{R}^{n \times m}$ of each 3D scan from the unlabeled pool the prediction $P_\theta(y|x)$ under the current model θ is computed, where n is the image height and m is the image width. For each pixel, the distance of the prediction to a maximum uncertainty value α is computed. The closer the predicted probability is to α, the more uncertain the model. Then, the uncertainty of a slice x is given as the negative sum of the class- and pixel-wise distance measures:

$$x^*_{LC}(x) = -\sum_{c=1}^{C} \sum_{i=1}^{n} \sum_{j=1}^{m} |\alpha - P_\theta(y_{c,i,j}|x)|$$

with

$$\alpha = \begin{cases} 0.5 & \text{in multi-label scenario} \\ \frac{1}{C} & \text{in single-label scenario with } C = \#classes \end{cases}$$

Entropy-based sampling Another approach to measure uncertainty is Shannon entropy. In this approach, the uncertainty of a slice $x \in \mathbb{R}^{n \times m}$ is calculated as the sum of the pixel-wise distance measures multiplied by the logarithm of the pixel-wise distance measures:

$$x^*_H(x) = \sum_{c=1}^{C} \sum_{i=1}^{n} \sum_{j=1}^{m} |\alpha - P_\theta(y_{c,i,j}|x)| \times \log_e |\alpha - P_\theta(y_{c,i,j}|x)|$$

with

$$\alpha = \begin{cases} 0.5 & \text{in multi-label scenario} \\ \frac{1}{C} & \text{in single-label scenario with } C = \#classes \end{cases}$$

Since the uncertainty of a model is usually similar for adjacent slices of a 3D scan, uncertainty sampling can result in multiple redundant slices of the same 3D scan being sampled. To mitigate this effect, we limit uncertainty sampling to one slice per 3D scan in each active learning iteration unless the number of 3D scans is less than the number of slices to be sampled.

Representativeness Sampling

Representativeness sampling aims to increase the diversity and representativeness of the sampled training set. In this work, we study one model-independent representativeness sampling strategy, stratified random sampling, and two model-driven strategies, clustering-based sampling and distance-based sampling.

Stratified random sampling In this most basic type of representativeness sampling, slices are sampled randomly from the unlabeled 3D scans, while ensuring that the same number of slices is sampled from each 3D scan (each 3D scan is considered a stratum).

Model-driven strategies measure image similarity using model-generated feature vectors that represent the content of an image slice in a compressed form. To obtain such feature vectors, for each unlabeled slice $x \in \mathbb{R}^{n \times m}$ of each 3D scan, the prediction under the current model θ is computed, and a feature vector is retrieved from one of the inner model layers. In our implementation, the bottleneck layer of the U-Net architecture is used to retrieve feature vectors. To reduce the dimensionality of the obtained feature vectors, spatial max-pooling is applied, followed by a principal component analysis (PCA). In our implementation, the spatial dimension is aggregated into a single value by spatial max-pooling, and then the feature channels are compressed to ten principal components.

Clustering-based representativeness sampling In the clustering-based approach, we apply mean shift clustering with a fixed bandwidth of 5 to the feature vectors. Unlabeled slices are selected from the obtained clusters such that each cluster is equally represented in the training set relative to its size.

Distance-based representativeness sampling In the distance-based approach, the average Euclidean distance between each feature vector of the set of unlabeled slices and each feature vector of the training set is computed:

$$s(x) = \frac{1}{N} \sum_{i=1}^{N} d(f_i, f_x)$$

where N is the training set size, f_i is the feature vector of the i-th slice in the training set, f_x is the feature vector of an unlabeled slice x, and d is the Euclidean distance. The unlabeled slices with the largest average distance to the feature vectors from the training set are selected as the most representative. A similar sampling technique is described in the work of Smailagic et al. (2018).

Strided Sampling

To combine the computational efficiency of a 2D model with the three-dimensional structure of medical image data, we propose a strided sampling strategy that generates pseudo-labels through interpolation. Since adjacent slices of a 3D scan contain redundant information, strided sampling extracts sub-blocks of 3D scans, consisting of a top, a bottom, and intermediate 2D slices, and selects only the top and the bottom slice for annotation. The intermediate slices are either not included in the training set, or pseudo-labels are generated by interpolating between the labels of the top and the bottom slice. The first case, strided sampling without interpolation, is equivalent to the equal interval query strategy proposed in the work of Zhang et al. (2019b). The second case, strided sampling with pseudo-labels, is similar to the works of Gorriz et al. (2017) and Zheng et al. (2020), who also make use of pseudo-labels. In our case, however, the pseudo-labels are not generated by a model but by interpolation, meaning that their quality is independent of the quality of the model. We experiment with two different interpolation methods, namely signed distance interpolation and morphological contour interpolation.

Signed distance interpolation This interpolation method computes distance maps for the top and bottom slices of a selected block and is inspired by a Stackoverflow question.[2] These distance maps are calculated separately for each class channel and indicate the signed distance to the edge of the segmentation mask for each pixel $x_{i,j}$ in a two-dimensional image slice $x \in \mathbb{R}^{n \times m}$. Pixels belonging to the foreground class are assigned a positive distance value and pixels belonging to the background class a negative distance value. To calculate the distance maps, the signed distance function $d(x)$ is used, which is defined as

$$d(x) = edt(x) - edt(x^{-1}) + 0.5$$

where x^{-1} is the inverted segmentation mask and

$$edt(x)_{i,j} = \|x_{i,j} - b_{i,j}\|$$

is the Euclidian distance transform where $b_{i,j}$ is the background pixel with the smallest Euclidean distance to input pixel $x_{i,j}$.

To obtain an interpolated segmentation mask, we interpolate between the distance maps of the top and bottom slices. All pixels with a positive distance value in the interpolated distance map are assigned to the foreground class and all other pixels to the background class.

Morphological contour interpolation For the morphological contour interpolation, we use an implementation of the Insight Toolkit (ITK) introduced by Zukić et al. (2016). The approach is based on decomposing a many-to-many correspondence into three fundamental cases: one-to-one, one-to-many, and zero-to-one correspondences (Albu et al., 2008).

---

2. *Interpolation between two images:* stackoverflow.com/questions/48818373/interpolate-between-two-images
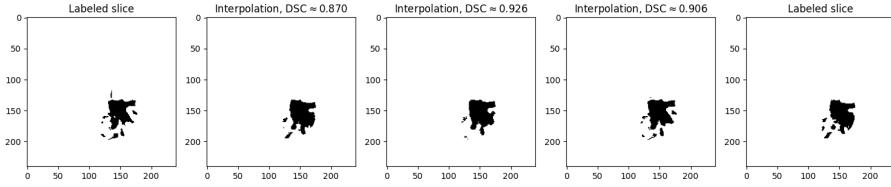
Figure 4: Example interpolation with signed distance method of a block of a brain tumor segmentation mask with block size $= 5$.
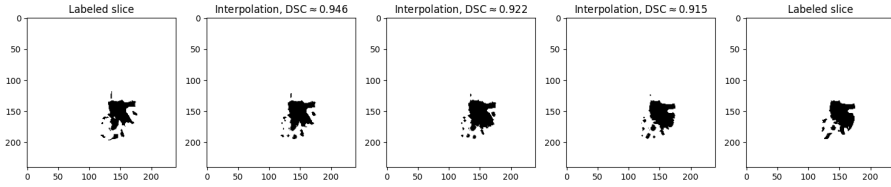
Figure 5: Example interpolation with morphological contour method of a block of a brain tumor segmentation mask with block size $= 5$.

If the segmentation masks of the top and bottom slice of a selected block do not overlap, no interpolation of the segmentation masks is possible. In such cases, no pseudo-labels are created and only the upper and lower slices of the block are added to the training set.

The sampling of blocks in strided sampling can be done either randomly, or according to another criterion, e.g. uncertainty. The distance between the top and bottom slice of a block is determined by the hyperparameter l, the maximum block size. If labels or pseudo-labels already exist for some slices of a selected block, we discard it and sample another, completely unlabeled block instead. If there are no more fully unlabeled blocks for the given block size l, the block size is reduced before block sampling continues. Once all slices in the dataset have been assigned either a label or a pseudo-label, the block size is reset to the original block size l, and block selection subsequently also accepts blocks with pseudo-labels, so that pseudo-labels are gradually replaced by true labels.

Appendix B.

To compare different query strategies, we developed an active learning simulation framework that allows evaluating active learning strategies on different medical image datasets. The framework provides two model training modes: classic training on a fully labeled dataset and active learning. Both modes are compatible with two-dimensional segmentation models, and the classic mode additionally supports three-dimensional models. The framework comes with eleven ready-to-use datasets, and custom datasets can be integrated with little effort. The provided datasets include the two-dimensional Breast Cancer Semantic Segmentation dataset (Amgad et al., 2019) and ten three-dimensional datasets from the Medical Segmentation Decathlon challenge (Antonelli et al., 2021). In the active learning mode, the framework implements the query strategies described in Section 3.1. Since the framework focuses on benchmarking and comparing different query strategies, it provides a reproducibility mode that enables different active learning strategies to be executed under the same conditions. For this purpose, all random processes, such as model and training pool initialization, are seeded, and only deterministic operations are used. In addition, the framework allows tracking various metrics and losses such as the Dice score, Hausdorff distance, sensitivity, and specificity. Several loss functions are implemented, including Dice loss (Milletari et al., 2016), cross-entropy loss, and focal loss (Lin et al., 2017). The framework is implemented using PyTorch[3] (Paszke et al., 2019) and PyTorch Lightning (Falcon and The PyTorch Lightning team, 2019).[4] Due to its modular structure, it can be easily extended with additional datasets, model architectures, query strategies, and evaluation metrics. More details can be found in the framework repository on GitHub[5] and in the framework documentation.[6]

---

3. *PyTorch framework:* `https://pytorch.org`

4. *PyTorch Lightning framework:* `https://www.pytorchlightning.ai`

5. *Active Segmentation framework:* `https://github.com/HealthML/active-segmentation`

6. *Active Segmentation framework documentation:* `https://healthml.github.io/active-segmentation/`

Appendix C.

In this supplementary material, we provide the mean validation Dice scores of the foreground classes for all tested active learning strategies for selected active learning iterations. The following tables show the average and standard deviation of the mean validation Dice scores from three experiment runs.

Table 1: Performance of single slice query strategies on the heart dataset.

Query Strategy	Active Learning Iteration				
	5	10	15	20	50
Random sampling	0.807 ± 0.049	0.868 ± 0.011	0.870 ± 0.006	0.887 ± 0.004	0.893 ± 0.004
Least confidence-based uncertainty sampling	0.835 ± 0.012	0.878 ± 0.007	0.887 ± 0.004	0.883 ± 0.010	0.897 ± 0.003
Entropy-based uncertainty sampling	0.839 ± 0.017	0.867 ± 0.012	0.881 ± 0.008	0.868 ± 0.018	0.887 ± 0.003
Stratified random sampling	0.802 ± 0.045	0.865 ± 0.008	0.872 ± 0.012	0.883 ± 0.004	0.899 ± 0.001
Clustering-based representativeness sampling	0.839 ± 0.021	0.869 ± 0.014	0.883 ± 0.004	0.886 ± 0.011	0.895 ± 0.005
Distance-based representativeness sampling	0.786 ± 0.009	0.813 ± 0.032	0.861 ± 0.024	0.881 ± 0.002	0.894 ± 0.008

Table 2: Performance of single slice query strategies on the hippocampus dataset.

Query Strategy	Active Learning Iteration				
	5	10	15	20	50
Random sampling	0.766 ± 0.016	0.803 ± 0.010	0.812 ± 0.012	0.819 ± 0.007	0.834 ± 0.005
Least confidence-based uncertainty sampling	0.747 ± 0.004	0.778 ± 0.005	0.784 ± 0.002	0.792 ± 0.005	0.816 ± 0.003
Entropy-based uncertainty sampling	0.759 ± 0.005	0.790 ± 0.006	0.798 ± 0.006	0.806 ± 0.003	0.825 ± 0.003
Stratified random sampling	0.777 ± 0.016	0.809 ± 0.007	0.818 ± 0.009	0.821 ± 0.004	0.835 ± 0.001
Clustering-based representativeness sampling	0.778 ± 0.013	0.799 ± 0.006	0.813 ± 0.004	0.817 ± 0.007	0.834 ± 0.003
Distance-based representativeness sampling	0.756 ± 0.003	0.799 ± 0.012	0.811 ± 0.011	0.815 ± 0.006	0.829 ± 0.005

Table 3: Performance of single slice query strategies on the prostate dataset.

Query Strategy	Active Learning Iteration				
	5	10	15	20	50
Random sampling	0.382 ± 0.103	0.500 ± 0.008	0.501 ± 0.025	0.524 ± 0.027	0.554 ± 0.013
Least confidence-based uncertainty sampling	0.425 ± 0.120	0.501 ± 0.042	0.498 ± 0.010	0.539 ± 0.007	0.563 ± 0.017
Entropy-based uncertainty sampling	0.418 ± 0.121	0.524 ± 0.025	0.559 ± 0.012	0.553 ± 0.021	0.579 ± 0.032
Stratified random sampling	0.409 ± 0.109	0.518 ± 0.010	0.535 ± 0.009	0.527 ± 0.032	0.549 ± 0.033
Clustering-based representativeness sampling	0.355 ± 0.086	0.448 ± 0.064	0.510 ± 0.036	0.515 ± 0.046	0.561 ± 0.026
Distance-based representativeness sampling	0.387 ± 0.091	0.505 ± 0.017	0.490 ± 0.018	0.543 ± 0.008	0.584 ± 0.029

For the strided sampling strategy, we experimented with two approaches for block selection, namely random sampling and entropy-based uncertainty sampling. For both approaches, we tested two interpolation methods, signed distance interpolation and morphological contour interpolation. In Section 4 and Section 5, we focus on the results obtained with random block selection and signed distance interpolation, as signed distance interpolation outperformed morphological contour interpolation (Table 10) and random sampling produced decent results on all datasets. Uncertainty-based strided sampling performed worse than the random sampling baseline on the hippocampus dataset, while in some cases, e.g., on the heart dataset, it performed slightly better. In Section 4 and Section 5, the term "strided sampling" always refers to random sampling-based strided sampling with signed distance interpolation. In the following, we provide the results of all tested variants.

Table 4: Performance of random sampling-based strided sampling strategies on the heart dataset.

Query Strategy	Active Learning Iteration				
	5	10	15	20	50
Random sampling	0.807 ± 0.049	0.868 ± 0.011	0.870 ± 0.006	0.887 ± 0.004	0.893 ± 0.004
No interpolation, block size 5	0.787 ± 0.038	0.855 ± 0.018	0.877 ± 0.016	0.886 ± 0.011	0.899 ± 0.004
Signed distance interpolation, block size 5	0.806 ± 0.025	0.864 ± 0.003	0.870 ± 0.013	0.890 ± 0.006	0.894 ± 0.005
Morphological contour interpolation, block size 5	0.766 ± 0.126	0.86 ± 0.02	0.877 ± 0.012	0.890 ± 0.014	0.893 ± 0.008
No interpolation, block size 10	0.804 ± 0.023	0.868 ± 0.008	0.889 ± 0.006	0.885 ± 0.014	0.888 ± 0.003
Signed distance interpolation, block size 10	0.820 ± 0.005	0.875 ± 0.006	0.888 ± 0.004	0.890 ± 0.005	0.896 ± 0.005
Morphological contour interpolation, block size 10	0.824 ± 0.029	0.867 ± 0.006	0.885 ± 0.006	0.889 ± 0.011	0.893 ± 0.007
No interpolation, block size 15	0.816 ± 0.039	0.877 ± 0.008	0.887 ± 0.004	0.886 ± 0.009	0.888 ± 0.004
Signed distance interpolation, block size 15	0.819 ± 0.016	0.870 ± 0.004	0.879 ± 0.009	0.885 ± 0.001	0.898 ± 0.002
Morphological contour interpolation, block size 15	0.810 ± 0.027	0.864 ± 0.013	0.871 ± 0.003	0.881 ± 0.001	0.893 ± 0.008

Table 5: Performance of random sampling-based strided sampling strategies on the hippocampus dataset. For the hippocampus dataset, block size 15 was not tested since all scans in this dataset contain less than 15 slices.

Query Strategy	Active Learning Iteration				
	5	10	15	20	50
Random sampling	0.766 ± 0.016	0.803 ± 0.010	0.812 ± 0.012	0.819 ± 0.007	0.834 ± 0.005
No interpolation, block size 5	0.775 ± 0.021	0.807 ± 0.008	0.817 ± 0.007	0.823 ± 0.004	0.838
Signed distance interpolation, block size 5	0.774 ± 0.014	0.808 ± 0.006	0.813 ± 0.010	0.818 ± 0.005	0.831 ± 0.001
Morphological contour interpolation, block size 5	0.762 ± 0.020	0.795 ± 0.008	0.801 ± 0.012	0.809 ± 0.010	0.811 ± 0.011
No interpolation, block size 10	0.785 ± 0.011	0.815 ± 0.001	0.821 ± 0.003	0.828 ± 0.001	0.836 ± 0.003
Signed distance interpolation, block size 10	0.786 ± 0.011	0.812 ± 0.001	0.818 ± 0.002	0.821 ± 0.005	0.829 ± 0.004
Morphological contour interpolation, block size 10	0.786 ± 0.011	0.810 ± 0.002	0.819 ± 0.003	0.823 ± 0.004	0.828

Table 6: Performance of random sampling-based strided sampling strategies on the prostate dataset.

Query Strategy	Active Learning Iteration				
	5	10	15	20	50
Random sampling	0.382 ± 0.103	0.500 ± 0.008	0.501 ± 0.025	0.524 ± 0.027	0.554 ± 0.013
No interpolation, block size 5	0.428 ± 0.131	0.502 ± 0.019	0.527 ± 0.020	0.538 ± 0.021	0.563 ± 0.004
Signed distance interpolation, block size 5	0.443 ± 0.073	0.508 ± 0.039	0.541 ± 0.012	0.507 ± 0.077	0.574 ± 0.026
Morphological contour interpolation, block size 5	0.456 ± 0.089	0.539 ± 0.019	0.519 ± 0.036	0.522 ± 0.022	0.582 ± 0.014
No interpolation, block size 10	0.394 ± 0.105	0.484 ± 0.041	0.512 ± 0.008	0.523 ± 0.028	0.580 ± 0.005
Signed distance interpolation, block size 10	0.398 ± 0.107	0.496 ± 0.015	0.541 ± 0.023	0.555 ± 0.041	0.543 ± 0.008
Morphological contour interpolation, block size 10	0.394 ± 0.102	0.518 ± 0.017	0.550 ± 0.017	0.538 ± 0.013	0.580 ± 0.002
No interpolation, block size 15	0.303 ± 0.164	0.429 ± 0.065	0.521 ± 0.025	0.550 ± 0.019	0.564 ± 0.019
Signed distance interpolation, block size 15	0.305 ± 0.136	0.394 ± 0.080	0.516 ± 0.019	0.488 ± 0.015	0.541 ± 0.011
Morphological contour interpolation, block size 15	0.269 ± 0.217	0.386 ± 0.087	0.491 ± 0.015	0.460 ± 0.031	0.552 ± 0.030

Table 7: Performance of uncertainty sampling-based strided sampling strategies on the heart dataset.

Query Strategy	Active Learning Iteration				
	5	10	15	20	50
Random sampling	0.807 ± 0.049	0.868 ± 0.011	0.870 ± 0.006	0.887 ± 0.004	0.893 ± 0.004
No interpolation, block size 5	0.834 ± 0.022	0.871 ± 0.004	0.886 ± 0.009	0.883 ± 0.020	0.892 ± 0.006
Signed distance interpolation, block size 5	0.807 ± 0.017	0.866 ± 0.011	0.889 ± 0.011	0.882 ± 0.009	0.893 ± 0.003
Morphological contour interpolation, block size 5	0.802 ± 0.002	0.868 ± 0.009	$0.89 \ \pm 0.01$	0.894 ± 0.005	0.893 ± 0.008
No interpolation, block size 10	0.851 ± 0.013	0.879 ± 0.002	0.884 ± 0.010	0.875 ± 0.014	0.892 ± 0.005
Signed distance interpolation, block size 10	0.822 ± 0.021	0.878 ± 0.008	0.878 ± 0.014	0.890 ± 0.007	0.894 ± 0.007
Morphological contour interpolation, block size 10	0.835 ± 0.013	0.876 ± 0.004	0.885 ± 0.002	0.890 ± 0.006	0.891 ± 0.008
No interpolation, block size 15	0.847 ± 0.011	0.879 ± 0.003	0.879 ± 0.006	0.878 ± 0.005	0.887 ± 0.002
Signed distance interpolation, block size 15	0.851 ± 0.004	0.868 ± 0.029	0.879 ± 0.006	0.879 ± 0.011	0.901 ± 0.003
Morphological contour interpolation, block size 15	0.835 ± 0.012	0.852 ± 0.010	0.860 ± 0.006	0.877 ± 0.008	0.884 ± 0.006

Table 8: Performance of uncertainty sampling-based strided sampling strategies on the hippocampus dataset. For the hippocampus dataset, block size 15 was not tested since all scans in this dataset contain less than 15 slices.

Query Strategy	Active Learning Iteration				
	5	10	15	20	50
Random sampling	0.766 ± 0.016	0.803 ± 0.010	0.812 ± 0.012	0.819 ± 0.007	0.834 ± 0.005
No interpolation, block size 5	0.750 ± 0.023	0.786 ± 0.009	0.800 ± 0.007	0.807 ± 0.005	0.828 ± 0.005
Signed distance interpolation, block size 5	0.747 ± 0.014	0.780 ± 0.006	0.794 ± 0.007	0.805 ± 0.007	0.822 ± 0.003
Morphological contour interpolation, block size 5	0.728 ± 0.024	0.770 ± 0.004	0.782 ± 0.011	0.783 ± 0.009	0.798 ± 0.003
No interpolation, block size 10	0.758 ± 0.012	0.790 ± 0.017	0.806 ± 0.008	0.813 ± 0.006	0.832 ± 0.003
Signed distance interpolation, block size 10	0.753 ± 0.003	0.790 ± 0.007	0.799 ± 0.007	0.802 ± 0.002	0.807 ± 0.014
Morphological contour interpolation, block size 10	0.747 ± 0.012	0.786 ± 0.013	0.791 ± 0.018	0.805 ± 0.010	0.807 ± 0.022

Table 9: Performance of uncertainty sampling-based strided sampling strategies on the prostate dataset.

Query Strategy	Active Learning Iteration				
	5	10	15	20	50
Random sampling	0.382 ± 0.103	0.500 ± 0.008	0.501 ± 0.025	0.524 ± 0.027	0.554 ± 0.013
No interpolation, block size 5	0.405 ± 0.100	0.499 ± 0.010	0.508 ± 0.015	0.538 ± 0.012	0.557 ± 0.013
Signed distance interpolation, block size 5	0.397 ± 0.098	0.514 ± 0.037	0.529 ± 0.031	0.530 ± 0.018	0.556 ± 0.054
Morphological contour interpolation, block size 5	0.402 ± 0.113	0.507 ± 0.031	0.534 ± 0.049	0.528 ± 0.011	0.551 ± 0.025
No interpolation, block size 10	0.384 ± 0.124	0.496 ± 0.006	0.550 ± 0.031	0.515 ± 0.054	0.571 ± 0.018
Signed distance interpolation, block size 10	0.406 ± 0.107	0.475 ± 0.078	0.560 ± 0.032	0.528 ± 0.018	0.540 ± 0.027
Morphological contour interpolation, block size 10	0.400 ± 0.112	0.507 ± 0.009	0.555 ± 0.028	0.519 ± 0.021	0.545 ± 0.023
No interpolation, block size 15	0.309 ± 0.165	0.443 ± 0.093	0.524 ± 0.025	0.507 ± 0.007	0.575 ± 0.012
Signed distance interpolation, block size 15	0.296 ± 0.129	0.414 ± 0.105	0.525 ± 0.010	0.547 ± 0.035	0.561 ± 0.035
Morphological contour interpolation, block size 15	0.301 ± 0.087	0.464 ± 0.079	0.506 ± 0.006	0.536 ± 0.021	0.558 ± 0.020

Table 10: Quality of pseudo-labels generated through interpolation in the strided sampling strategy. The Dice scores provided in this table were calculated using the ground-truth labels.

Dataset	Interpolation Approach	Max. Block Size	Dice Score
Heart	signed distance	5	0.974 ± 0.045
Heart	signed distance	10	0.950 ± 0.065
Heart	signed distance	15	0.930 ± 0.085
Heart	morphological contour	5	0.969 ± 0.050
Heart	morphological contour	10	0.948 ± 0.066
Heart	morphological contour	15	0.928 ± 0.082
Hippocampus	signed distance	5	0.937 ± 0.102
Hippocampus	signed distance	10	0.917 ± 0.176
Hippocampus	morphological contour	5	0.914 ± 0.147
Hippocampus	morphological contour	10	0.903 ± 0.206
Prostate	signed distance	5	0.879 ± 0.119
Prostate	signed distance	10	0.847 ± 0.147
Prostate	signed distance	15	0.811 ± 0.167
Prostate	morphological contour	5	0.877 ± 0.124
Prostate	morphological contour	10	0.847 ± 0.146
Prostate	morphological contour	15	0.809 ± 0.168