IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 4, 2021, accepted December 18, 2021, date of publication January 6, 2022, date of current version January 13, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3141021

Robust Segmentation Models Using an Uncertainty Slice Sampling-Based Annotation Workflow

GRZEGORZ CHLEBUS[1,2], **ANDREA SCHENK**[1], **HORST K. HAHN**[1,3], **BRAM VAN GINNEKEN**[1,2], **AND HANS MEINE**[1,4]

[1]Fraunhofer Institute for Digital Medicine MEVIS, 28359 Bremen, Germany
[2]Diagnostic Image Analysis Group, Department of Medical Imaging, Radboud University Medical Center, 6525 Nijmegen, The Netherlands
[3]Department of Computer Science and Electrical Engineering, Jacobs University, 28759 Bremen, Germany
[4]Medical Image Computing Group, University of Bremen, 28359 Bremen, Germany

Corresponding author: Grzegorz Chlebus (grzegorz.chlebus@mevis.fraunhofer.de)

This work was supported by the Fraunhofer-Gesellschaft.

ABSTRACT Semantic segmentation neural networks require pixel-level annotations in large quantities to achieve a good performance. In the medical domain, such annotations are expensive because they are time-consuming and require expert knowledge. Active learning optimizes the annotation effort by devising strategies to select cases for labeling that are the most informative to the model. In this work, we propose an uncertainty slice sampling (USS) strategy for the semantic segmentation of 3D medical volumes that selects 2D image slices for annotation and we compare it with various other strategies. We demonstrate the efficiency of USS on a CT liver segmentation task using multisite data. After five iterations, the training data resulting from USS consisted of 2410 slices (4% of all slices in the data pool) compared to 8121 (13%), 8641 (14%), and 3730 (6%) slices for uncertainty volume (UVS), random volume (RVS), and random slice (RSS) sampling, respectively. Despite being trained on the smallest amount of data, the model based on the USS strategy evaluated on 234 test volumes significantly outperformed models trained according to the UVS, RVS, and RSS strategies and achieved a mean Dice index of 0.964, a relative volume error of 4.2%, a mean surface distance of 1.35 mm, and a Hausdorff distance of 23.4 mm. This was only slightly inferior to 0.967, 3.8%, 1.18 mm, and 22.9 mm achieved by a model trained on all available data. Our robustness analysis using the 5th percentile of Dice and the 95th percentile of the remaining metrics demonstrated that USS not only resulted in the most robust model compared to other strategies, but also outperformed the model trained on all data according to the 5th percentile of Dice (0.946 vs. 0.945) and the 95th percentile of mean surface distance (1.92 mm vs. 2.03 mm).

INDEX TERMS Active learning, convolutional neural network, deep learning, segmentation, uncertainty sampling.

I. INTRODUCTION

Semantic segmentation of medical images plays a key role in many treatment planning workflows. In recent years, segmentation algorithms utilizing deep neural networks have provided state-of-the-art results for many segmentation tasks [1]–[3]. Training such systems typically requires large datasets with pixel-level annotations to achieve good performance. In the medical domain, such annotations require expert knowledge and are time-consuming and thus expensive to obtain. Although many pre-trained segmentation models are currently available, it is known that neural networks underperform when applied to data coming from different sites or imaging protocols [4], [5]. Therefore, training on the annotated target data is recommended to achieve optimal performance.

Various approaches have been developed to optimize the annotation effort. Çiçek *et al.* [6] proposed a semi-automated segmentation setup in which a user only needs to annotate several image slices to obtain dense segmentation.

The associate editor coordinating the review of this manuscript and approving it for publication was Kumaradevan Punithakumar.

O'Neil *et al.* [7] investigated the utility of crowdsourcing from non-expert annotators for medical tasks and concluded that an acceptable quality ground truth can be obtained given a sufficient number of observers per scan. Cai *et al.* [8] proposed a weakly supervised 3D segmentation approach, where a CNN model is trained using dense annotations returned by the GrabCut method initialized with RECIST strokes. The CNN model was initially trained on 2D annotations, and then propagated segmentation to neighboring slices in an iterative manner. Bortsova *et al.* [9] presented a semi-supervised learning approach that trains an FCN model using unlabeled data by exploiting the fact that the segmentations of deformed versions of the same image should be consistent. Leveraging unlabeled data improved model performance significantly when compared with a supervised learning setting. Tajbakhsh *et al.* [10] investigated transfer learning in this context and showed that pre-trained models fine-tuned on a target domain are not only outperforming but are also less sensitive to the size of the target training set than models trained from scratch.

Active learning is a prominent technique that primarily focuses on annotation effort optimization [11]. Its setup consists of query strategies that select cases from a pool of unlabeled examples for annotation and an oracle that provides labels for the selected cases. Typically, the query strategies aim to maximize the model performance while maintaining a low annotation cost. The query strategies usually take the model uncertainty [12], [13], sample representativeness [14], or both of them [15], [16] into account to guide the selection process. Smailagic *et al.* [16] proposed a strategy for a classification problem that uses predictive entropy to identify the most uncertain examples in a data pool. Among these examples, those that maximize the distance to the training set in a feature space, defined by the output of one of the model layers, were selected. The addition of the distance maximization criterion resulted in a 32% reduction in the number of required labeled examples. A similar approach to a segmentation task was demonstrated by Yang *et al.* [15], where an ensemble was used to select uncertain cases defined as those with the highest average per-pixel output variance. Additionally, examples with the highest representativeness according to the cosine similarity distance based on descriptors obtained from the model were selected for manual annotation. Models trained using this approach achieved a state-of-the-art performance using only 50% of the training data. Wang *et al.* [12] introduced cost-effective active learning for image classification, where in addition to manually labeled uncertain cases, those with high confidence were pseudo-labeled by the model and added to the training set with no annotation effort. Sinha *et al.* [17] proposed a variational adversarial query strategy that employs a discriminator network to differentiate between labeled and unlabeled examples. This approach selects points that are not well represented in the labeled set without the need for explicit uncertainty measurements on the main task.

While most of the pool-based active learning research focuses on uncertainty quantification and example representativeness, little attention has been paid to the incorporation of partial annotations into the workflow. Partial annotations are particularly relevant to semantic segmentation tasks because they can result in a substantial reduction in the annotation effort [6], especially when dealing with the volumetric data. In this work, we propose an uncertainty slice sampling (USS) query strategy to optimize the annotation workflow of 3D medical images. Our strategy selects 2D image slices for annotation from a pool of unlabeled 3D volumes using predictive entropy as the uncertainty measure. We see this as a way to increase the variability of the training set without the need for explicit modelling of example representativeness. We demonstrate the efficiency of our strategy on a CT liver segmentation task using multisite data. We analyze the proposed strategy together with several alternatives: uncertainty volume sampling (UVS), random volume sampling (RVS), and random slice sampling (RSS). The primary goal of this study is to provide an extensive comparative evaluation of the proposed USS query strategy.

II. METHODS

A. NEURAL NETWORK ARCHITECTURE

We used a 3D anisotropic u-net (au-net) architecture (see Fig. 1), which is a modified version of the commonly used encoder-decoder u-net segmentation network design [18]. The model works on five resolution levels. In the upper two levels, convolutional layers work only along x and y spatial dimensions and the remaining levels contain separable 3D convolutions to minimize the number of trainable parameters. Each convolution layer is followed by a batch normalization layer [19] and the ReLU activation function. Max pooling (transposed 3D convolution) layers are used in the encoder (decoder) for transitioning down (up) between the resolution levels. A dropout layer [20] with a drop rate $p = 0.25$ is placed at each resolution level in the decoder path to prevent overfitting and facilitate Monte Carlo (MC) sampling used for the uncertainty estimation.

B. DATA PREPROCESSING

For training, all CT volumes were rescaled into Hounsfield units and resampled to a 1.0 mm × 1.0 mm × 1.5 mm voxel size.

C. UNCERTAINTY ESTIMATION

Several uncertainty estimation methods for segmentation models have been proposed, including the volume variation coefficient [21], prediction variance [15], predictive entropy [16], [21]–[23], and mutual information [22]. These methods require multiple samples, which are typically obtained from an ensemble or via MC dropout [24]. In our work, we used the predictive entropy as, to the best of our knowledge, it is one of the most frequently used uncertainty measures and can be seen as a de facto standard.

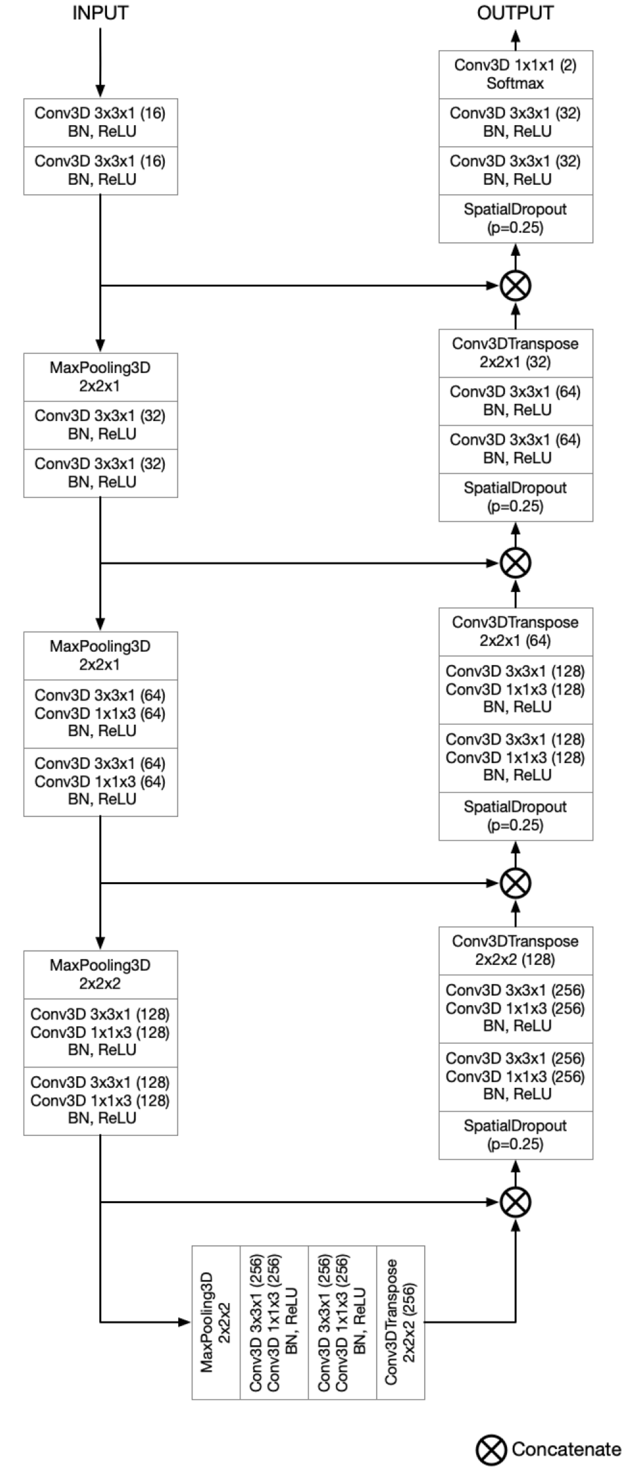

INPUT
OUTPUT
Conv3D 3x3x1 (16) BN, ReLU
Conv3D 3x3x1 (16) BN, ReLU
MaxPooling3D 2x2x1
Conv3D 3x3x1 (32) BN, ReLU
Conv3D 3x3x1 (32) BN, ReLU
MaxPooling3D 2x2x1
Conv3D 3x3x1 (64) Conv3D 1x1x3 (64) BN, ReLU
Conv3D 3x3x1 (64) Conv3D 1x1x3 (64) BN, ReLU
MaxPooling3D 2x2x2
Conv3D 3x3x1 (128) Conv3D 1x1x3 (128) BN, ReLU
Conv3D 3x3x1 (128) Conv3D 1x1x3 (128) BN, ReLU
MaxPooling3D 2x2x2
Conv3D 3x3x1 (256) Conv3D 1x1x3 (256) BN, ReLU
Conv3D 3x3x1 (256) Conv3D 1x1x3 (256) BN, ReLU
Conv3DTranspose 2x2x2 (256)
Conv3DTranspose 2x2x2 (128)
Conv3D 3x3x1 (256) Conv3D 1x1x3 (256) BN, ReLU
Conv3D 3x3x1 (256) Conv3D 1x1x3 (256) BN, ReLU
SpatialDropout (p=0.25)
Conv3DTranspose 2x2x1 (64)
Conv3D 3x3x1 (128) Conv3D 1x1x3 (128) BN, ReLU
Conv3D 3x3x1 (128) Conv3D 1x1x3 (128) BN, ReLU
SpatialDropout (p=0.25)
Conv3DTranspose 2x2x1 (32)
Conv3D 3x3x1 (64) BN, ReLU
Conv3D 3x3x1 (64) BN, ReLU
SpatialDropout (p=0.25)
Conv3D 3x3x1 (32) BN, ReLU
Conv3D 3x3x1 (32) BN, ReLU
SpatialDropout (p=0.25)
Conv3D 1x1x1 (2) Softmax
⊗ Concatenate

FIGURE 1. **Anisotropic u-net (au-net) architecture. The model has 4 975 346 trainable parameters.**

The predictive entropy for a voxel $\mathbf{x}$ is computed as follows:

$$U(\mathbf{x}) = -\sum_{c}\left(\frac{1}{n}\sum_{n} p_n(y=c|\mathbf{x})\right)\ln\left(\frac{1}{n}\sum_{n} p_n(y=c|\mathbf{x})\right) \tag{1}$$

where c is the number of classes, n is the number of samples, and $p_n(y = c|\mathbf{x})$ is the softmax probability of input $\mathbf{x}$ belonging to class c in the nth sample. In our experiments, we obtained $n = 20$ samples via MC dropout to capture the epistemic uncertainty accounting for the uncertainty in the model parameters [25]. We found 20 samples to make a reasonable trade-off between the uncertainty resolution and the computation speed.

Volume-level uncertainty Similar to [23], we used the average voxel-wise predictive entropy as a measure of uncertainty at the volume level. As CT volumes in our dataset have a substantial slice count variability, computing the average over all voxels would cause the volume-level uncertainty scores for CTs with a small slice count (e.g., abdominal acquisitions) to be overestimated when compared to volumes with a larger slice count (e.g., whole body acquisitions). To account for this, we excluded voxels outside of a dilated liver mask ($11 \times 11 \times 11$ box kernel) from the computation.

Slice-level uncertainty is computed as the average of the predictive entropy for all voxels belonging to a given slice. As the slice size variability across our dataset is negligible, there is no need to exclude any voxels from the computation as in the case of volume-level uncertainty.

D. QUERY STRATEGIES

Query strategies select which samples from the unlabeled data pool should be annotated and added to the training set. In our study, we investigated two orthogonal dimensions of query strategies: i) selection of 3D image volumes vs. 2D image slices and ii) random vs. uncertainty-based sampling. These dimensions define four sampling strategies that work as follows.

Uncertainty Volume Sampling (UVS) selects volumes with the largest volume-level uncertainty.

Random Volume Sampling (RVS) selects volumes at random.

Uncertainty Slice Sampling (USS) selects slices with the largest slice-level uncertainty from a set of slice candidates coming from all volumes. The slice candidates of one volume correspond to the local maxima of a slice-level uncertainty profile (slice-level uncertainty vs. slice index curve, see Fig. 2). To avoid selecting neighboring slices as candidates, we require the minimum distance between the local maxima to be at least five slices.

Random Slice Sampling (RSS) selects slices at random.

E. TRAINING SETUP

All models were trained with a mini-batch size of 2 using $180 \times 180 \times 4$ image patches that were padded (reflect mode) on each side with 92 voxels along x and y, and 20 voxels along the z spatial dimension to account for valid convolutions. The optimization was performed using the Adam optimizer with a 10^{-5} learning rate. All models were applied to the validation data every 1000 training steps and the best model according to the Jaccard index was used for the final evaluation.

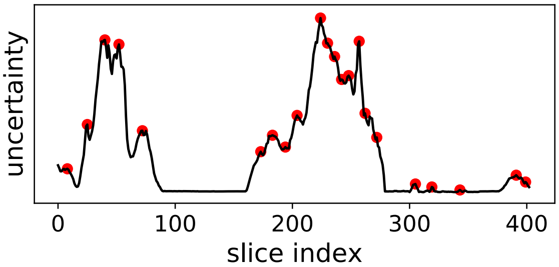


FIGURE 2. Slice-level uncertainty profile of one volume. The red dots correspond to slice candidates available to the uncertainty slice sampling strategy.

The minimal slice-related output patch size of our model along the z-dimension is four. For slice-sampling strategies (USS and RSS), where single slices are annotated, the loss needs to be computed only on labeled slices. To enable this, we used a weighted version of the soft dice loss [26]:

$$L_{DSC} = 1 - \frac{2\sum_i w_i y_i p_i}{\sum_i w_i y_i + \sum_i w_i p_i} \tag{2}$$

where y_i is ith voxel label (1 - liver, 0 - background), p_i is the output probability of the liver class, w_i is 1 for annotated voxels and 0 otherwise, and i runs over all voxels in a mini-batch.

Stratified patch sampling was employed to speed up the training by ensuring that at least one patch in a mini-batch contains liver pixels.

F. EVALUATION METRICS

We evaluated the segmentation quality with four commonly used metrics: Dice index (DICE), relative volume error (RVE), mean surface distance (MSD), and Hausdorff distance (HD) [27]. These metrics are defined as follows:

$$\mathrm{DICE}(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|} \tag{3}$$

$$\mathrm{RVE}(X, Y) = \frac{|V_X - V_Y|}{V_Y} \cdot 100\% \tag{4}$$

$$\begin{aligned}\mathrm{MSD}(X, Y) = \frac{1}{|S_X| + |S_Y|}\Bigg(&\sum_{x \in S_X} \min_{y \in S_Y} d(x, y) \\ &+ \sum_{y \in S_Y} \min_{x \in S_X} d(x, y)\Bigg)\end{aligned} \tag{5}$$

$$\mathrm{HD}(X, Y) = max\{\sup_{x \in S_X} \inf_{y \in S_Y} d(x, y), \sup_{y \in S_Y} \inf_{x \in S_X} d(x, y)\} \tag{6}$$

where X and Y are sets of voxels belonging to the test and reference object, respectively, S_X is the set of surface voxels of X, V_X denotes the volume of X, and $d(x, y)$ is the Euclidean distance between points x and y.

Evaluation results on the test set are reported using the mean and the 5th and 95th percentiles. We used these percentile-based measures to assess the impact of investigated query strategies on models' robustness.

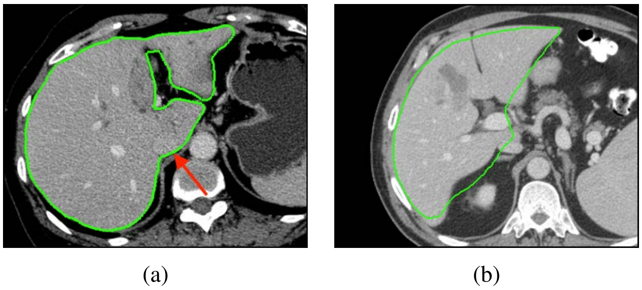

(a) (b)

FIGURE 3. Examples of excluded ***LiTS*** cases: (a) inconsistent annotations - inferior vena cava included in the liver mask (b) poor segmentation quality.

III. DATA

In this work, we simulated the active learning-based annotation workflow by using abdominal CT volumes with reference liver segmentations. To increase the statistical power of our study, we decided to use five sets of imaging data containing 484 CT volumes in total. Three of these datasets are proprietary and were obtained via cooperation with our clinical partners: Yokohama City University, Yokohama, Japan (*Yokohama*), Städtisches Klinikum Dresden, Dresden, Germany (*Dresden*), and Radboud University Clinical Center, Nijmegen, the Netherlands (*Rumc*). The remaining two datasets (*LiTS*, *CHAOS*) come from publicly organized challenges [3], [28]. Reference liver segmentations for the proprietary data were created manually by experienced clinical experts using dedicated annotation software [29]. The challenge data comes with training liver masks that we used as reference segmentations. We found 65 out of 131 *LiTS* cases to have either a poor reference segmentation quality or inconsistent annotations (e.g., inferior vena cava included in the liver mask) with the rest of our data. We decided to exclude these cases from our dataset, as training on such cases could bias our study. See Fig. 3 for examples of excluded LiTS cases.

The resulting combined dataset was divided into three subsets containing 240, 10, and 234 cases that were used as data pool, validation set, and test set, respectively. The reader is referred to Tab. 1 for more details.

IV. EXPERIMENTS

Our goal was to compare the investigated query strategies by keeping the required annotation time at a comparable level. We made the following assumptions on the annotation time.

1) The annotation time required for a slice without a liver is negligible in comparison to the time needed for a slice with a liver.
2) The time required for the annotation of N slices with a liver coming from the same volume allows for the annotation of approximately $N/3$ slices with a liver from different volumes.

The second assumption was derived empirically from our observations that for a segmentation of one volume one needs to manually draw contours on average on every third slice

TABLE 1. Dataset details.

Dataset Name	#Volumes (pool/val/test)	Voxel Size [mm]	Resolution
Yokohama	218 (119/2/97)	[0.5-0.8]×[0.5-0.8]×[0.8-1.6]	512×512×[170-376]
Rumc	100 (46/2/52)	[0.6-1.0]×[0.6-1.0]×[0.7-1.5]	512×512×[378-1057]
Dresden	80 (36/2/42)	[0.6-0.9]×[0.6-0.9]×[0.5-5.0]	512×512×[71-1160]
LiTS	66 (31/2/33)	[0.6-1.0]×[0.6-1.0]×[0.7-5.0]	512×512×[74-987]
CHAOS	20 (8/2/10)	[0.6-0.8]×[0.6-0.8]×[1.0-2.0]	512×512×[81-266]

when using modern segmentation tools with interpolation functionality.

A. QUERY STRATEGY COMPARISON

To evaluate the efficiency of the query strategies, we performed five active learning iterations and compared the models after each iteration. A model trained on five fully annotated cases was used as the initial model. In each iteration, we trained a model from scratch and used the best model according to the validation score for the evaluation and uncertainty estimation. All models were trained for a maximum of 30 epochs to guarantee that the model training and selection of cases for annotation could be performed overnight.

For the volume sampling strategies (UVS and RVS), we added five volumes to the training set in each iteration. We denote the average count of liver slices added in each UVS iteration by $\tilde{N}_S^{\text{liver}}$. Based on our assumption on the annotation process, in the USS strategy we added $N_S = \tilde{N}_S^{\text{liver}}/3$ slices in each iteration, whereas in the RSS we sampled random slices until N_S liver slices were selected.

B. CONVERGED MODELS COMPARISON

To evaluate the representativeness of the training sets obtained with the investigated query strategies after five active learning iterations, we compared the models trained on these training sets until convergence. This was motivated by the fact that 30 epochs were not sufficient for some models to converge.

V. RESULTS

A. ANNOTATION EFFORT

The initial model was trained using five fully annotated volumes, accounting for 1410 (501) annotated slices (liver slices). On average, 1342 (576) and 1446 (558) slices were annotated in each UVS and RVS iteration, respectively. In the USS and RSS iterations, we sampled $N_S = 200$ ($576/3 \approx 200$) slices. Consequently, in the 5th and final iteration, the models were trained using 13% (13%), 14% (12%), 4% (5%), and 6% (6%) of all slices (liver slices) available in the data pool for the UVS, RVS, USS, and RSS strategy, respectively. Further details can be found in Tab. 2.

B. QUERY STRATEGY COMPARISON

For all strategies, an overall improvement in all segmentation metrics was observed over the course of the active learning iterations (see Fig. 4). The mean performance metrics of the initial model were 0.925, 7.87%, 2.94 mm, and 36.3 mm for DICE, RVE, MSD, and HD, respectively. The mean performance metrics of the models resulting from the fifth (last) iteration of the investigated query strategies were in the ranges of [0.956, 0.960], [3.98, 6.04]%, [1.4, 1.73] mm, and [25.6, 28.6] mm for DICE, RVE, MSD, and HD, respectively (see Tab. 4-7). The model obtained from the last iteration of the USS strategy was significantly better (Wilcoxon signed-rank test) than the models resulting from the remaining strategies. When considering the segmentation metrics for all five iterations, the USS strategy resulted in the best-performing model most of the time. Interestingly, this strategy produced the worst model according to all metrics in the first iteration, even failing to improve the performance of the initial model for RVE and HD. None of the models trained with 4%-14% of the data was able to match the results of the model trained on the whole data pool that achieved on average the values of 0.967, 3.8%, 1.18 mm and 22.9 mm for DICE, RVE, MSD and HD, respectively.

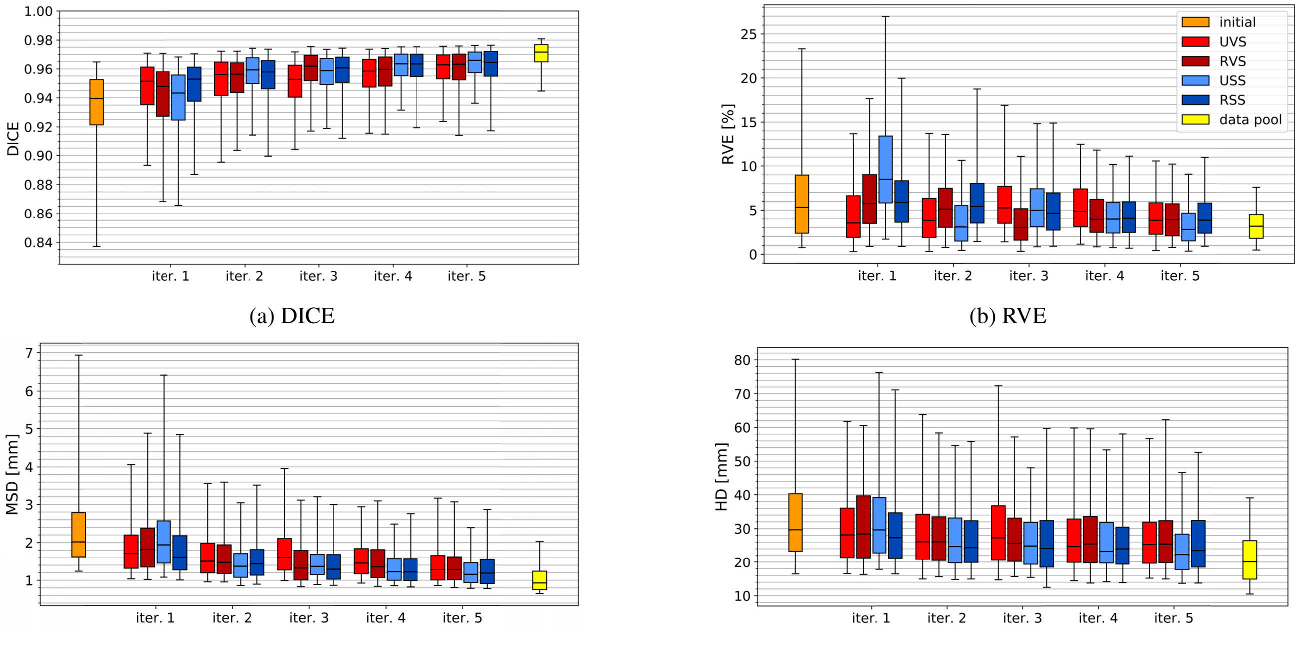
TABLE 2. Training data summary (annotated slices / annotated liver slices / unique patients) over the course of five active learning iterations for each investigated query strategy.

	UVS	RVS	USS	RSS
initial	1410 / 501 / 5			
iter. 1	2541 / 1045 / 10	3206 / 1070 / 10	1610 / 669 / 72	1873 / 701 / 195
iter. 2	3651 / 1537 / 15	4631 / 1608 / 15	1810 / 797 / 104	2323 / 901 / 233
iter. 3	5105 / 2224 / 20	5813 / 2150 / 20	2010 / 947 / 125	2790 / 1101 / 236
iter. 4	6941 / 2761 / 25	7094 / 2687 / 25	2210 / 1123 / 149	3278 / 1301 / 239
iter. 5	8121 / 3382 / 30	8641 / 3292 / 30	2410 / 1292 / 160	3730 / 1501 / 240
data pool	60513 / 26400 / 240			

TABLE 3. Training step counts for models used for evaluation of four investigated query strategies over five active learning iterations. For reference, data for the initial and data pool models are given.

	UVS	RVS	USS	RSS
initial	12 000			
iter. 1	35 000	37 000	40 000	46 000
iter. 2	61 000	64 000	61 000	67 000
iter. 3	80 000	83 000	71 000	94 000
iter. 4	75 000	85 000	91 000	117 000
iter. 5	119 000	98 000	98 000	139 000
iter. 5 (converged)	356 000	297 000	291 000	465 000
data pool	825 000			

For volume sampling strategies, no clear difference between uncertainty-based and random sampling was



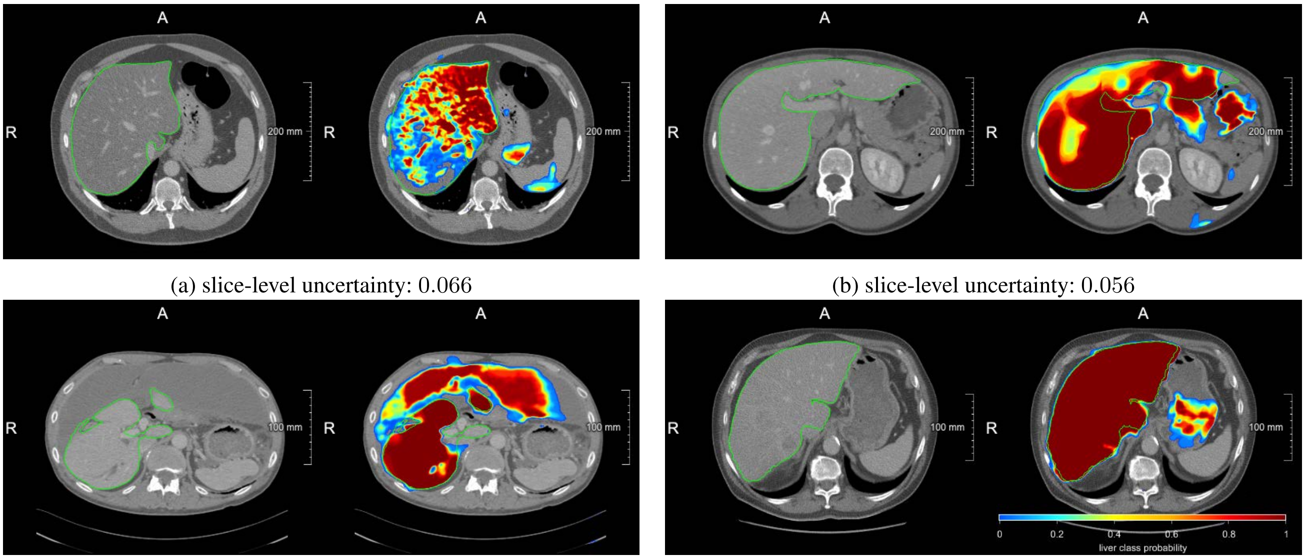
(a) DICE

(b) RVE

(c) MSD

(d) HD

FIGURE 4. Box plots summarizing evaluation results for models trained throughout five active learning iterations. For reference, results of the initial and whole data pool models are included. Lower (upper) caps correspond to the 5th (95th) percentile.



(a) slice-level uncertainty: 0.066

(b) slice-level uncertainty: 0.056

(c) slice-level uncertainty: 0.052

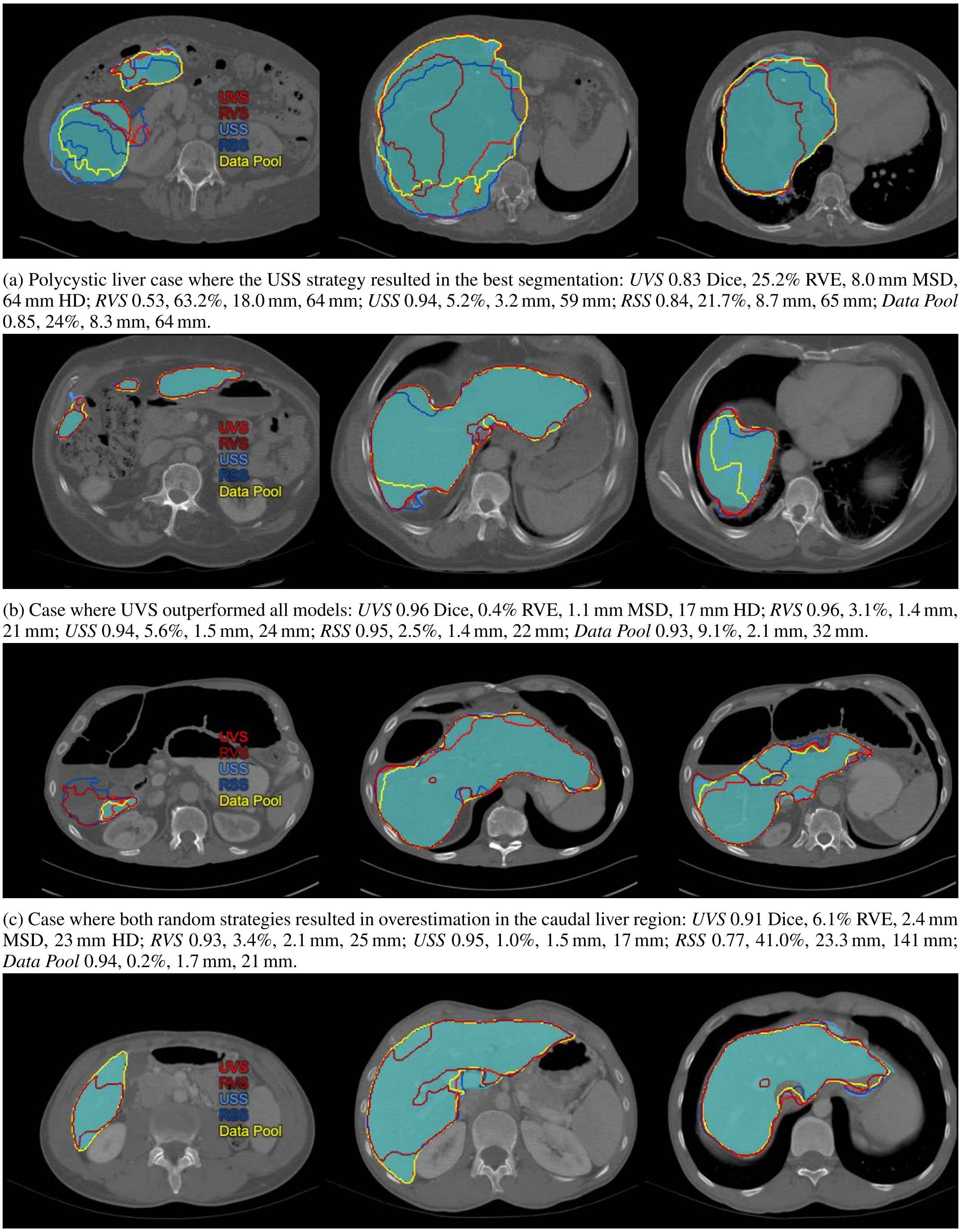
(d) slice-level uncertainty: 0.020

FIGURE 5. Sample slices selected in the first USS iteration with overlaid liver reference segmentation (green contour) and model liver probability output (heatmap): (a)-(c) slices with the biggest slice level uncertainty, (d) slice with the lowest uncertainty among selected ones.

observed. In the case of slice sampling strategies, the uncertainty-based sampling resulted in an improvement in all the metrics (excluding the first iteration).

C. CONVERGED MODELS COMPARISON

Training until convergence using data resulting from the fifth active learning iteration resulted in an improvement in all the segmentation metrics for all the query strategies, with the exception of the RVE metric for the USS strategy (see Tab. 8 and Fig. 7). The model trained on data resulting from the RSS strategy required the longest training time (465 000 training steps) among the models trained using the investigated query strategies, whereas the USS model needed the shortest (291 000). For reference, the model trained on the whole data pool needed 825 000 steps until convergence (see Tab. 3 for more details). Among the investigated strategies, the USS model had the best performance when compared to the other models, with most of the differences being significant. None of the models achieved comparable results to the whole data pool model, but we observed that the USS



(a) Polycystic liver case where the USS strategy resulted in the best segmentation: *UVS* 0.83 Dice, 25.2% RVE, 8.0 mm MSD, 64 mm HD; *RVS* 0.53, 63.2%, 18.0 mm, 64 mm; *USS* 0.94, 5.2%, 3.2 mm, 59 mm; *RSS* 0.84, 21.7%, 8.7 mm, 65 mm; *Data Pool* 0.85, 24%, 8.3 mm, 64 mm.

(b) Case where UVS outperformed all models: *UVS* 0.96 Dice, 0.4% RVE, 1.1 mm MSD, 17 mm HD; *RVS* 0.96, 3.1%, 1.4 mm, 21 mm; *USS* 0.94, 5.6%, 1.5 mm, 24 mm; *RSS* 0.95, 2.5%, 1.4 mm, 22 mm; *Data Pool* 0.93, 9.1%, 2.1 mm, 32 mm.

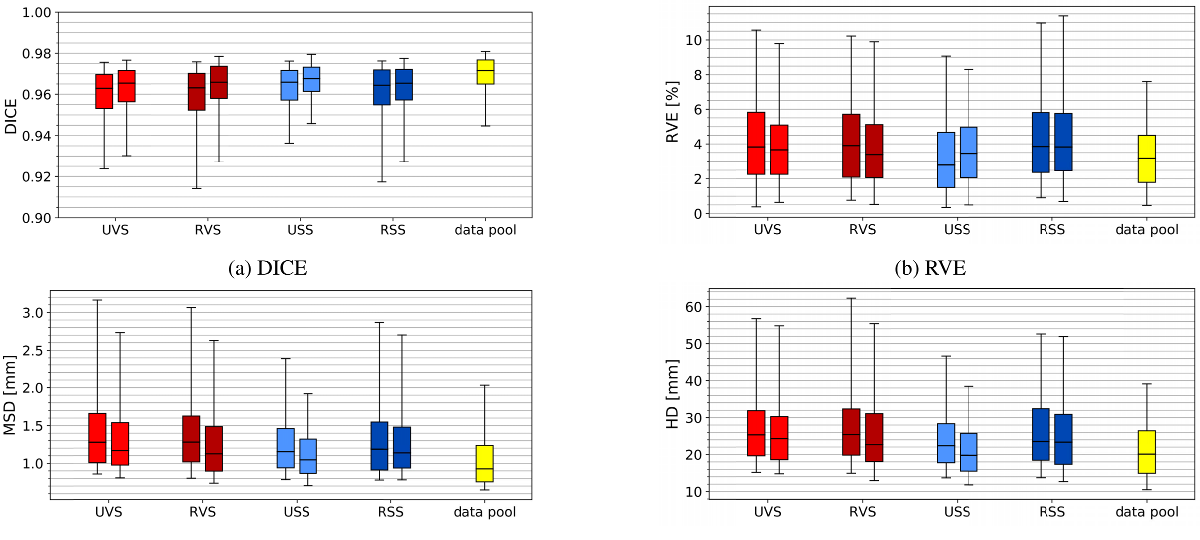
(c) Case where both random strategies resulted in overestimation in the caudal liver region: *UVS* 0.91 Dice, 6.1% RVE, 2.4 mm MSD, 23 mm HD; *RVS* 0.93, 3.4%, 2.1 mm, 25 mm; *USS* 0.95, 1.0%, 1.5 mm, 17 mm; *RSS* 0.77, 41.0%, 23.3 mm, 141 mm; *Data Pool* 0.94, 0.2%, 1.7 mm, 21 mm.

(d) Case for which all strategies except RVS achieved very good segmentation performance: *UVS* 0.97 Dice, 0.1% RVE, 0.9 mm MSD, 18 mm HD; *RVS* 0.91, 11.8%, 2.9 mm, 36 mm; *USS* 0.96, 1.8%, 1.0 mm, 20 mm; *RSS* 0.97, 1.6%, 0.9 mm, 20 mm; *Data Pool* 0.97, 0.0%, 0.8 mm, 20 mm.

FIGURE 6. Representative examples presenting segmentation output of the converged models and the model trained on the whole data pool.

model resulted in more robust segmentations than the whole data pool model with respect to the 5th percentile for DICE and the 95th percentile for the MSD and HD metrics.

To quantify an impact of uncertainty-based vs. random sampling and slice vs. volume sampling on the 5th percentile of DICE and 95th percentile of the remaining metrics, we performed quantile regression analysis using the evaluation results of the converged models (Tab. 9). The uncertainty sampling resulted in a significant improvement over the random sampling expressed by an increase in the 5th percentile of DICE by 0.012 and a decrease in the 95th percentile of MSD by 0.41. Similarly, changing from the



(a) DICE

(b) RVE

(c) MSD

(d) HD

FIGURE 7. Box plots summarizing evaluation results for models trained for maximum of 30 epochs (left) and until convergence (right) using data from the fifth iteration. For reference, results of the whole data pool model are included. Lower (upper) caps correspond to the 5th (95th) percentile.

TABLE 4. DICE for models trained using the investigated query strategies over five active learning iterations. Reported values represent mean and 5th to 95th percentiles (in the parentheses).

	UVS	RVS	USS	RSS
initial	0.925 (0.837, 0.965)			
iter. 1	0.942 (0.893, 0.971)	0.937 (0.868, 0.971)***	0.932 (0.866, 0.968)***	**0.943 (0.887, 0.97)**
iter. 2	0.947 (0.895, 0.972)***	0.947 (0.903, 0.972)***	**0.951 (0.914, 0.974)**	0.948 (0.899, 0.974)***
iter. 3	0.945 (0.904, 0.972)***	**0.955 (0.917, 0.975)**	0.953 (0.919, 0.973)***	0.951 (0.912, 0.974)*
iter. 4	0.952 (0.916, 0.974)***	0.952 (0.915, 0.974)***	**0.958 (0.932, 0.975)**	0.957 (0.919, 0.975)
iter. 5	0.956 (0.924, 0.976)***	0.956 (0.914, 0.976)***	**0.96 (0.936, 0.976)**	0.956 (0.917, 0.976)***
data pool	0.967 (0.945, 0.981)			

Best result in a row according to the mean is indicated in bold. Asterisks mark significantly worse results when compared to the best result (one-sided Wilcoxon signed-rank test): $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

TABLE 5. RVE for models trained using the investigated query strategies over five active learning iterations. Reported values represent mean and 5th to 95th percentiles (in the parentheses).

	UVS	RVS	USS	RSS
initial	7.87 (0.72, 23.32)			
iter. 1	**5.58 (0.27, 13.66)**	6.94 (0.85, 17.64)***	11.62 (1.71, 26.97)***	8.05 (0.85, 19.96)***
iter. 2	**5.1 (0.3, 13.69)**	6.15 (0.73, 13.57)***	5.51 (0.42, 10.63)	7.7 (1.42, 18.74)***
iter. 3	7.14 (1.4, 16.88)***	**4.19 (0.31, 11.08)**	6.16 (0.83, 14.8)***	7.32 (0.91, 14.89)***
iter. 4	6.06 (1.14, 12.44)***	**5.12 (0.83, 11.79)**	5.19 (0.72, 10.17)	5.25 (0.68, 11.11)
iter. 5	5.1 (0.38, 10.56)***	4.81 (0.76, 10.22)***	**3.98 (0.34, 9.07)**	6.04 (0.9, 10.97)***
data pool	3.8 (0.46, 7.6)			

Best result in a row according to the mean is indicated in bold. Asterisks mark significantly worse results when compared to the best result (one-sided Wilcoxon signed-rank test): $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

TABLE 6. MSD for models trained using the investigated query strategies over five active learning iterations. Reported values represent mean and 5th to 95th percentiles (in the parentheses).

	UVS	RVS	USS	RSS
initial	2.94 (1.24, 6.94)			
iter. 1	**2.1 (1.03, 4.06)**	2.17 (1.02, 4.88)**	2.81 (1.08, 6.41)***	2.28 (1.0, 4.84)
iter. 2	**1.85 (0.96, 3.56)**	1.86 (0.95, 3.6)	1.96 (0.86, 3.04)	2.04 (0.89, 3.5)
iter. 3	2.12 (0.99, 3.96)***	**1.64 (0.83, 3.11)**	1.64 (0.89, 3.2)***	2.05 (0.86, 3.0)
iter. 4	1.69 (0.92, 2.94)***	1.68 (0.84, 3.09)***	**1.52 (0.86, 2.48)**	1.58 (0.82, 2.75)
iter. 5	1.59 (0.86, 3.16)***	1.58 (0.8, 3.06)***	**1.4 (0.78, 2.39)**	1.73 (0.78, 2.87)**
data pool	1.18 (0.65, 2.03)			

Best result in a row according to the mean is indicated in bold. Asterisks mark significantly worse results when compared to the best result (one-sided Wilcoxon signed-rank test): $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

TABLE 7. HD for models trained using the investigated query strategies over five active learning iterations. Reported values represent mean and 5th to 95th percentiles (in the parentheses).

	UVS	RVS	USS	RSS
initial	36.3 (16.4, 80.2)			
iter. 1	**32.1 (16.5, 61.8)**	32.1 (16.3, 60.5)	36.4 (17.8, 76.3)*	33.2 (16.4, 71.1)
iter. 2	29.9 (14.9, 63.8)	29.9 (15.6, 58.3)	**29.6 (14.8, 54.6)**	29.6 (14.9, 55.8)
iter. 3	32.4 (14.7, 72.3)***	29.5 (15.6, 57.1)	**27.5 (15.4, 48.0)**	29.5 (12.5, 59.7)
iter. 4	28.3 (14.4, 59.8)	28.7 (13.7, 59.5)	**27.4 (14.1, 53.3)**	27.9 (13.9, 58.0)
iter. 5	28.5 (15.2, 56.7)***	28.6 (14.9, 62.3)***	**25.6 (13.7, 46.6)**	28.2 (13.7, 52.6)***
data pool	22.9 (10.5, 39.1)			

Best result in a row according to the mean is indicated in bold. Asterisks mark significantly worse results when compared to the best result (one-sided Wilcoxon signed-rank test): $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

TABLE 8. Evaluation results for the converged models. Reported values represent mean and 5th to 95th percentiles (in the parentheses).

	DICE	RVE [%]	MSD [mm]	HD [mm]
UVS	0.959 (0.930, 0.977)***	4.67 (0.65, 9.78)**	1.45 (0.81, 2.73)***	27.0 (14.8, 54.8)***
RVS	0.959 (0.927, 0.978)***	4.41 (0.53, 9.89)	1.44 (0.74, 2.63)***	26.4 (12.9, 55.4)***
USS	**0.964 (0.946, 0.979)**	**4.20 (0.49, 8.30)**	**1.35 (0.71, 1.92)**	**23.4 (11.8, 38.4)**
RSS	0.959 (0.927, 0.977)***	5.00 (0.69, 11.38)***	1.54 (0.78, 2.70)***	27.1 (12.7, 52.0)***
data pool	0.967 (0.945, 0.981)	3.80 (0.47, 7.60)	1.18 (0.65, 2.03)	22.9 (10.5, 39.1)

Best result for given metric according to the mean is indicated in bold (model trained on the whole data pool is excluded from comparison). Asterisks mark significantly worse results when compared to the best result (one-sided Wilcoxon signed-rank test): $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

TABLE 9. Quantile regression results quantifying an impact of uncertainty vs. random sampling and slice vs. volume sampling on the 5th percentile of DICE and the 95th percentile of RVE, MSD, and HD. This analysis was done using the evaluation results of the converged models.

	DICE†		RVE [%]‡		MSD [mm]‡		HD [mm]‡	
	coefficient	p-value	coefficient	p-value	coefficient	p-value	coefficient	p-value
Intercept	0.920 [0.914, 0.925]	<0.001	11.21 [9.26, 13.16]	<0.001	2.90 [2.64, 3.16]	<0.001	59.6 [50.0, 69.3]	<0.001
Uncertainty Sampling	0.012 [0.005, 0.019]	<0.001	-1.90 [-4.21, 0.42]	0.1	-0.41 [-0.73, -0.08]	0.015	-4.8 [-16.3, 6.6]	0.4
Slice Sampling	0.012 [0.005, 0.019]	<0.001	-0.80 [-3.12, 1.51]	0.5	-0.48 [-0.81, -0.16]	0.004	-7.7 [-19.1, 3.7]	0.2

Results from 0.05(†), 0.95(‡) quantile regression. Values in brackets denote 95% confidence interval.

volume to slice sampling resulted in an increase in the 5th percentile DICE by 0.012 and a decrease in the 95th percentile of MSD by 0.48. The effect on the remaining metrics did not pass the significance test (95% confidence interval included 0).

VI. CONCLUSION AND DISCUSSION

In this work, we proposed an uncertainty slice sampling (USS) strategy in the context of pool-based active learning. Our strategy selects 2D image slices from a pool of 3D volumes using aggregated voxel-wise predictive entropy as the uncertainty measure. We evaluated the proposed strategy on a CT liver segmentation task and compared it with random slice sampling (RSS), uncertainty volume sampling (UVS), and random volume sampling (RVS) strategies. The model trained by using the USS data (4% of the available data) achieved significantly better results than the models resulting from the remaining strategies. Although after five active learning iterations the USS model was performance-wise inferior in mean performance to the model trained on all the available data, it provided a more robust segmentation output as measured by the 5th percentile of DICE and the 95th percentile of MSD. We hypothesize that this can be attributed to the differences in the training set composition. The training set resulting from the USS contains a larger proportion of difficult and rare cases compared to the whole data-pool training set, which effectively makes the model see them more frequently during the training process. Fig. 6 shows sample outputs of the investigated models including two hard cases from the test set: a polycystic (Fig. 6a) and a resected (Fig. 6b) liver. We think that the robustness of the whole data pool model could be increased by employing hard example mining during training to dynamically adjust the sampling rate of difficult examples [30], [31].

Selecting only uncertain cases in the course of active learning can overload the model with difficult examples causing a performance drop. This was observed for the USS strategy after the first iteration when the model performed substantially worse than its random counterpart (see Fig. 4).

The uncertainty information made a substantial contribution to the USS model performance. The model resulting from the RSS strategy, which selects slices at random, was inferior to the USS model in all but the initial iteration. Moreover, in the comparison of the converged models, the RSS model had the worst performance among all the investigated strategies. For the volume-level query strategies, we did not observe clear differences in results between UVS and RVS. A possible explanation for this is the uncertainty captured at the volume level that provides a low signal-to-noise ratio if a model makes only small local errors and produces accurate segmentations. We think that this was the case, as according to the DICE metric, which describes the global level segmentation quality, the initial model provided mostly correct outputs (0.925 average DICE). The final converged models managed to improve this metric by only ∼4%.

Because our proposed strategy relies on the model's uncertainty to query cases, the confidence calibration of a model can have a substantial impact on which cases are deemed uncertain. Recently, it has been shown that modern deep neural networks do not output well-calibrated probabilities and tend to be overconfident [32]. In our work, we used MC dropout to improve the calibration quality of models trained with the Dice loss [23]. Sample probability maps produced by our model are shown in Fig. 5. We think that an investigation of various calibration techniques, e.g., deep ensembles and temperature scaling, in the context of active learning could be an interesting future research direction.

In our study, we focused on a detailed investigation of the proposed query strategy, and we did not intend to achieve state-of-the-art results on the CT liver segmentation task. We are aware that in addition to using more training data, increasing the model capacity or changing its architecture could further boost the segmentation performance. Examining the impact of various neural network designs on the efficiency of our proposed query strategy would, in our opinion, lead to a compelling experiment. Moreover, we think that an investigation of active learning along with AutoML frameworks such as nnU-net [33], which make state-of-the-art segmentation accessible to people without ML expertise, would be an interesting extension of our work.

Limitations Our conclusions are based on the assumptions on the annotation effort derived from our experience of annotation workflows. These assumptions might not hold for some annotators or if different labeling tools are used.

ACKNOWLEDGMENT

The authors would like to thank our clinical partners from Yokohama City University, Yokohama, Japan, Städtisches Klinikum Dresden, Dresden, Germany, and Radboud University Clinical Center, Nijmegen, the Netherlands for providing the imaging data used in this study, also would like to thank the organizers of the LiTS and the CHAOS competitions for making the training data publicly available, and also would like to thank Christiane Engel and Andrea Koller for providing manual segmentations.

REFERENCES

[1] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, van B. Ginneken, and M. Bilello, "The medical segmentation decathlon," 2021, *arXiv:2106.05735*.

[2] H. Bogunović, F. Venhuizen, S. Klimscha, S. Apostolopoulos, A. Bab-Hadiashar, U. Bagci, M. F. Beg, L. Bekalo, Q. Chen, C. Ciller, and K. Gopinath, "RETOUCH: The retinal OCT fluid detection and segmentation benchmark and challenge," *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1858–1874, Aug. 2019.

[3] P. Bilic, P. F. Christ, E. Vorontsov, G. Chlebus, H. Chen, Q. Dou, C.-W. Fu, X. Han, P. A. Heng, J. Hesser, and S. Kadoury, "The liver tumor segmentation benchmark (LiTS)," 2019, *arXiv:1901.04056*.

[4] E. Gibson, Y. Hu, N. Ghavami, H. U. Ahmed, C. Moore, M. Emberton, H. J. Huisman, and D. C. Barratt, "Inter-site variability in prostate segmentation accuracy using deep learning," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 506–514.

[5] N. Karani, K. Chaitanya, C. Baumgartner, and E. Konukoglu, "A lifelong learning approach to brain mr segmentation across scanners and protocols," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 476–484.

[6] Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2016, pp. 424–432.

[7] A. Q. O'Neil, J. T. Murchison, E. J. V. Beek, and K. A. Goatman, "Crowdsourcing labels for pathological patterns in ct lung scans: Can non-experts contribute expert-quality ground truth? in *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. Cham, Switzerland: Springer, 2017, pp. 96–105.

[8] J. Cai, Y. Tang, L. Lu, A. P. Harrison, K. Yan, J. Xiao, L. Yang, and R. M. Summers, "Accurate weakly supervised deep lesion segmentation on CT scans: Self-paced 3D mask generation from RECIST," 2018, *arXiv:1801.08614*.

[9] G. Bortsova, F. Dubost, L. Hogeweg, I. Katramados, and M. D. Bruijne, "Semi-supervised medical image segmentation via learning consistency under transformations," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 810–818.

[10] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1299–1312, Mar. 2016.

[11] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *J. Artif. Intell. Res.*, vol. 4, pp. 129–145, Mar. 1996.

[12] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2591–2600, Dec. 2016.

[13] M. Gorriz, A. Carlier, E. Faure, and X. Giro-i-Nieto, "Cost-effective active learning for melanoma segmentation," 2017, *arXiv:1711.09168*.

[14] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," 2017, *arXiv:1708.00489*.

[15] L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen, "Suggestive annotation: A deep active learning framework for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2017, pp. 399–407.

[16] A. Smailagic, P. Costa, H. Young Noh, D. Walawalkar, K. Khandelwal, A. Galdran, M. Mirshekari, J. Fagert, S. Xu, P. Zhang, and A. Campilho, "MedAL: Accurate and robust deep active learning for medical image analysis," in *Proc. 17th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2018, pp. 481–488.

[17] S. Sinha, S. Ebrahimi, and T. Darrell, "Variational adversarial active learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5972–5981.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[20] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, Jun. 2014.

[21] A. G. Roy, S. Conjeti, N. Navab, and C. Wachinger, "Inherent brain segmentation quality control from fully convnet Monte Carlo sampling," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 664–672.

[22] J. Mukhoti and Y. Gal, "Evaluating Bayesian deep learning methods for semantic segmentation," 2018, *arXiv:1811.12709*.

[23] A. Mehrtash, W. M. Wells, C. M. Tempany, P. Abolmaesumi, and T. Kapur, "Confidence calibration and predictive uncertainty estimation for deep medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 12, pp. 3868–3878, Jul. 2020.

[24] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.

[25] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" 2017, *arXiv:1703.04977*.

[26] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2017, pp. 240–248.

[27] G. Chlebus, H. Meine, S. Thoduka, N. Abolmaali, B. van Ginneken, H. K. Hahn, and A. Schenk, "Reducing inter-observer variability and interaction time of MR liver volumetry by combining automatic CNN-based liver segmentation and manual corrections," *PLoS ONE*, vol. 14, no. 5, May 2019, Art. no. e0217228.

[28] A. E. Kavur, N. S. Gezer, M. Baris, S. Aslan, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkan, and B. Baydar, "CHAOS challenge–combined (CT-MR) healthy abdominal organ segmentation," *Med. Image Anal.*, vol. 69, Apr. 2021, Art. no. 101950.

[29] A. Schenk, G. P. Prause, and H.-O. Peitgen, "Local-cost computation for efficient segmentation of 3d objects with live wire," *Proc. SPIE*, vol. 4322, pp. 1357–1364, Jul. 2001.

[30] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 761–769.

[31] C. Bian, X. Yang, J. Ma, S. Zheng, Y.-A. Liu, R. Nezafat, P.-A. Heng, and Y. Zheng, "Pyramid network with online hard example mining for accurate left atrium segmentation," in *Proc. Int. Workshop Stat. Atlases Comput. Models Heart*. Cham, Switzerland: Springer, 2018, pp. 237–245.

[32] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1321–1330.

[33] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "NnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Dec. 2021.

GRZEGORZ CHLEBUS received the B.Sc. degree in biomedical engineering and the M.Sc. degree in electronics and telecommunications from the Wrocław University of Science and Technology, Wrocław, Poland, in 2012 and 2013, respectively. He is currently pursuing the Ph.D. degree in medical image analysis with Radboud University Medical Center, Nijmegen, The Netherlands. The M.Sc. thesis was written at the Computer Vision Lab Group, ETH Zurich, Zurich, Switzerland. He is currently a Scientist at Fraunhofer MEVIS, Bremen, Germany. He also works as a Senior Applied Scientist for Qualtrics, Kraków, Poland. His research interests include semantic segmentation, model uncertainty quantification, and anomaly detection.

ANDREA SCHENK received the Diploma degree in mathematics and computer science from Technical University Dortmund, in 1997, and the Ph.D. degree in medical engineering from the University of Bremen, in 2012. She studied in Dortmund, Germany, and Dublin, Ireland. From 1998 to 1999, she was a Research Assistant with the CeVis Institute, before she joined as a Senior Scientist with the Center for Medical Diagnostic Systems and Visualization (MeVis), in 2000, currently part of the Fraunhofer-Gesellschaft. She was a Guest Lecturer with the University of Bremen, Jacobs University Bremen, and the University of Applied Sciences, Bremerhaven. Since 2012, she has been the Head of Liver Research, and since 2015, she has been a member of the Management Board of Fraunhofer MEVIS, Bremen, Germany. She is the author of more than 110 articles. Her research interests include image processing of multimodal data, clinical decision support, and image guided therapies, with a focus on abdominal and oncology applications.

HORST K. HAHN received the Ph.D. degree *(summa cum laude)* in the field of quantitative medical imaging from the University of Bremen, in 2005. He studied physics with the minors mathematics, computer science, and physiology with the University of Bayreuth, Paul Sabatier University Toulouse, and Heidelberg University. In 2006, he became the Deputy Director of the non-profit MeVis Research GmbH and was involved in its transformation into the Fraunhofer Institute for Medical Image Computing (MEVIS), in January 2009, which he has been leads as the Institute Director, since October 2012. He was appointed as an Adjunct Professor in medical visualization, and four years later, a Professor in medical imaging with Jacobs University Bremen. In January 2012, he was a Visiting Professor with the Diagnostic Image Analysis Group, Radboud University Medical Center, Nijmegen. He is the author of eight patents and over 280 publications in the field of computer-assisted medicine. In Ph.D. degree, he received the Bruker Daltonik Special Award as part of the Bremen Study Prize, in 2005.

BRAM VAN GINNEKEN received the Ph.D. degree in computer-aided diagnosis in chest radiography from the Image Sciences Institute, in 2001. He studied physics at the Eindhoven University of Technology and Utrecht University. He is currently a Professor in medical image analysis with the Radboud University Medical Center, and the Chair of the Diagnostic Image Analysis Group. He also works for Fraunhofer MEVIS, Bremen, Germany. He is also the Founder of Thirona, a company that develops software and provides services for medical image analysis. He has coauthored over 200 publications in international journals. He is a member of the Fleischner Society and the Editorial Board of *Medical Image Analysis*. He pioneered the concept of challenges in medical image analysis.

HANS MEINE received the Ph.D. degree in computer science from the University of Hamburg, in 2008, for his fundamental research on image segmentation algorithms and data structures. He has always been interested in using his knowledge to solve actual problems in various domains and joined Fraunhofer MEVIS to focus on medical applications, in 2011. He is currently responsible for coordinating strategic developments in the fields of image analysis and deep learning across various projects. He is also teaching with the University of Bremen and giving courses in diverse, interdisciplinary contexts. His research interests include image segmentation and precise quantitative measurements from 2D or 3D images or videos and on uncertainty quantification.