Epistemic and aleatoric uncertainties reduction with rotation variation for medical image segmentation with ConvNets

Ge Zhang¹ · Hao Dang¹ · Yulong Xu¹

Received: 6 September 2021 / Accepted: 14 December 2021

Published online: 21 January 2022

© The Author(s) 2022

Abstract

The deep convolutional neural network (ConvNet) achieves significant segmentation performance on medical images of various modalities. However, the isolated errors in a large testing set with various tumor conditions are not acceptable in clinical practice. This is usually caused in inadequate training and noise inherent during data collection, which are recognized as epistemic and aleatoric uncertainties in deep learning-based approaches. In this paper, we analyze the two types of uncertainties in medical image segmentation tasks and propose a reduction method by training models with data augmentation. The shelter zones in images are reduced with 2D imaging on surfaces of different angles from 3D organs. Rotation transformation and noise are estimated by Monte Carlo simulation with prior parameter distributions, and the aleatoric uncertainty is quantized in this process. Experiments on segmentation of computed tomography images demonstrate that overconfident incorrect predictions are reduced through uncertainty reduction and that our method outperforms prediction baselines based on epistemic and aleatoric estimation.

Keywords Convolutional neural networks · Image segmentation · Epistemic · Aleatoric

1 Introduction

As an essential task in many surgical applications, e.g., diagnosis, treatment and recovery [1], image segmentation is applied in recognizing the boundaries of tumors in organs. Due to large variations in the segmentation target among patients, it is still far from accurate and reliable when applying segmentation to images from organic bodies [2]. The application is not usable in clinical practice even only isolated errors occur during segmentation due to the high precision requirement. Therefore, the reduction in uncertainties is critical in accurate and reliable segmentation. With quantization of epistemic and aleatoric uncertainties, we can assess the reliability of the results and guide human intervention when needed.

Segmentation methods based on convolutional neural networks (ConvNets) have investigated the uncertainty estimation issue in image classification and recognition [3–5]. However, the output always consists of high-level predictions, e.g., image annotations and network bounding box parameters. These methods do not provide uncertainty estimation on the pixelwise level, which is crucial in image segmentation [6]. Moreover, the training process requires a large quantity of labeled training data, which is difficult to obtain for medical images due to the specialization of annotation. Consequently, the training samples for deep learning-based approaches in medical image segmentation are comparatively small [7]. The limited training samples bring much uncertainty and, in turn, result in unreliable segmentation. One typical mechanism to solve the issue is

✉ Ge Zhang, zhangge@hactcm.edu.cn; Hao Dang, danglee@hactcm.edu.cn; Yulong Xu, flyxyl@126.com | ¹Henan University of Chinese Medicine, 156 Jinshui Road, Zhengzhou, China.

data augmentation technology [8], which involves flipping, cropping, rotating and scaling. However, there are not enough in-depth studies on the implementation of medical image segmentation.

Uncertainty estimation has been studied in deep neural networks for high-level image label prediction output, which can be categorized into two major types: epistemic and aleatoric. In the field of medical image segmentation, the former uncertainty is introduced in different sensors accuracy and label boundaries of organs or lesions, which can be overcome by using enough accurately labeled training data. The latter one depends on the amount of data lie outside of the training data, which can be reducible through an increase in measurement precision [9]. Recent works [10, 11] focus on test-time drop-out-based uncertainty during training, which can be recognized as epistemic uncertainty estimation. The aleatoric uncertainty is transferred to epistemic uncertainty through mapping input data with a unified Bayesian deep learning framework [12]. However, the proposed method has not been utilized for segmentation with medical data, especially 3D images.

Motivated by the above observations, we propose a data augmentation method for medical image segmentation by utilizing images captured from different angles. An image captured from an organ is recognized as the result of an acquisition process that contains geometric transforms, where the overlapped areas are identified in this process. Images captured from other angles of the same area are input into the training process to compensate for the overlapped area, which provides additional data for predicting the segmentation distribution. The prediction is achieved through Monte Carlo (MC) sampling on the predictive output based on the same trained neural network model. The augmentation can recover the miscaptured information and achieve better characterization during training and, in turn, predict better segmentation results.

The contributions of this paper can be summarized as threefold. First, we quantize the epistemic and aleatoric uncertainties that are accumulated during training and segmentation. Second, we propose a data augmentation method with training images of different rotation variations that facilitates uncertainty reduction, especially in medical image segmentation, which contains a small number of annotated training samples. The segmentation performance is improved without additional labeled training data, which are expensive to acquire. Third, the confidence level information of segmentation results can be calculated for human intervention purposes. The potential missegmented area is provided to the user for better diagnosis results. It is possible to reduce the workload of medical personnel significantly and avoid unnecessary misdiagnoses by machine learning only.

2 Related work

A Bayesian network estimates the epistemic uncertainty by learning the posterior distribution of weighted parameters in image segmentation. Due to the difficulty in implementation and high computational cost, drop-out at test time is cast as an approximation to represent the uncertainty [13]. Stochastic variational gradient descent (SVGD) is used in ConvNet [14] to estimate parameter uncertainty. However, it requires adopting multiple stochastic forward passes and calculating the average to obtain the uncertainty of weighted parameters. Moreover, the SVGD implementations focus on image classification and annotation. Other approaches, such as Markov chain Monte Carlo (MCMC) [15], Monte Carlo batch normalization (MCBN) [16] and variational Bayesian methods [17, 18], are also proposed to estimate the uncertainty. However, they have not been utilized in training networks for medical image segmentation. A scalable ensemble of multiple models for uncertainty estimation is proposed to learn the epistemic uncertainty [19], but the paper does not provide any solution to uncertainty measurement in medical image segmentation.

Neural networks have relatively high variance, which makes it difficult to replicate the results [25]. Bootstrap [26] and deep ensemble [19] methods are used to train multiple neural network models to reduce the variance. A summary of related studies conducted on uncertainty estimation using deep learning technologies is presented in Table 1. Every single network has its characteristics, and the predicted error is applied to epistemic or aleatoric uncertainty estimation. The average operation reduces the sensitivities from the training mechanism, parameter initialization and dataset details, which achieve better performance than any single model. A unified Bayesian deep learning network is proposed to map the input data to the uncertainty originating from variations in images for segmentation and classification [12]. However, their dataset used in experiments is natural images rather than medical images. As the first application to medical image analysis, superresolution diffusion MR brain images were utilized in uncertainty modeling with ConvNet in 2016 [27]. They provided an in-depth discussion on the accuracy improvement from predictive variance and MC sample variance uncertainty reduction. Later, a lung nodule detection system was proposed by leveraging MC sample variance [28]. They estimated the uncertainty in a two-stage network and achieved better segmentation results, but uncertainty measurements were not properly analyzed in this article. Recently, literature [29, 30] further discussed MC

Table 1 A summary of uncertainty estimation methods in deep learning pipelines for image classification, detection and segmentation

Algorithm	Reference	Uncertainty	Applied to	Limitations
Monte Carlo Dropout	[12]	Epistemic	Segmentation	High computational cost
variational Bayesian	[18]	Epistemic	Segmentation	High computational cost
Deep Ensemble	[19]	Epistemic	Classification	Multiple training
M Heads	[20]	Epistemic	Detection	Not for segmentation
SVGD	[14]	Epistemic	Classification	Parameter uncertainty only
Probabilistic U-Net	[21]	Epistemic	Segmentation	Softmax probability only
Hie. Prob. U-Net	[22]	Epistemic	Segmentation	Softmax probability only
PHiSeg	[23]	Epistemic	Segmentation	Do not capture the full posterior
Seg. Variability Estimation	[10]	Aleatoric	Segmentation	Do not capture the full posterior
Two-Stage Model Variable-Input	[24]	Aleatoric	Segmentation	Multiple initialization
Test-time Augmentation	[8]	Aleatoric	Segmentation	Low reliability and efficiency

Fig. 1 Flowchart of epistemic and aleatoric uncertainties reduction method with rotation variation for medical image segmentation with ConvNets

dropout uncertainty as a potential technique being used in medical image analysis, and learned a dynamic threshold through uncertainty estimation for high-confidence pseudolabels [31]. They achieved better performance on small object segmentation, but there was not any analysis of the necessity and impact of uncertainty estimation on medical image segmentation.

3 Materials and methods

The proposed method is described in detail in this section. In the analysis of epistemic and aleatoric uncertainties, we deduce the confidence of segmentation through a Bayesian neural network (BNN) [32] with Monte Carlo dropout. In the uncertainty reduction method, the training data are augmented to a larger size for better feature representation, such as tumors in human organs. Specifically, we utilize images at different rotation angles to enhance the annotated data in data augmentation, which is helpful in characterizing the overlapped areas and may result in better

segmentation performance. The overall flowchart of our method is illustrated in Fig. 1.

3.1 Epistemic uncertainty

The neural network model can be recognized as a conditional distribution model, which consists of an input distribution, an output prediction and a weight parameter. The prior distribution is used as the network parameter in BNN training. The output is a distribution of functions rather than a single function in our implementation. The confidence level of the result can be determined by consulting the distribution. The traditional method uses variational inference to calculate the posterior distribution. Although accurate, the training efficiency and computational complexity are significant in BNN training. We use MC dropout to deduce the Bayesian distribution and modify the loss function to determine the uncertainty.

The MC dropout samples the training data for limited iterations and generates an estimation of the posterior distribution. It is an approximate estimation of the posterior distribution that provides information on whether the input data exist in the learned distribution. This piece of information is regarded as epistemic uncertainty.

According to the deduction process in [33], dropout is used both at training and testing time as a random generator. This means that we ignore some of the neurons during training and testing. To obtain the expected model output and the confidence information, we simulate a network output with input y and treat the stochastic regularization technique (SRT) [34] as if it is being used during training. Assume the data are fed to the network T times and the average of the result is calculated as:

$$\mathbb{E}(y) = \frac{1}{T} \sum_t x_t, \quad (1)$$

where $\mathbb{E}(\mathbf{y})$ is the predictive mean of our approximate predictive distribution and the predictive variance (the expected uncertainty) can be calculated by performing variance operation on it:

$$\text{Var}(\mathbf{y}) = \sigma^2 + \frac{1}{T} \sum_{t=1}^T f(\mathbf{y}_t(\mathbf{x}))^T f(\mathbf{y}_t(\mathbf{x})) - \mathbb{E}[\mathbf{y}]^T \mathbb{E}[\mathbf{y}], \quad (2)$$

where $\mathbf{x}$ denotes the input features from training images and the predictive mean $\mathbb{E}[\mathbf{y}]$ denotes the expected model output given the input $\mathbf{x}$. This process is repeated T times for T independently identical distributions $\{\mathbf{y}_1(\mathbf{x}), \dots, \mathbf{y}_T(\mathbf{x})\}$. These output values are empirical samples from an approximate predictive distribution. σ^2 denotes the aleatoric uncertainty and will be discussed in the following section.

3.2 Aleatoric uncertainty

Unlike epistemic uncertainty reduction, aleatoric uncertainty captures noise inherent during the data collection. For segmentation tasks on 3D images, voxel-wise uncertainty estimation is more practical. The cost function in training can be deduced according to [33]:

$$L(\theta) = \frac{1}{N} \sum_{i=1}^N \frac{\|\mathbf{y}_i - f(\mathbf{x}_i)\|^2}{2\sigma(\mathbf{x}_i)^2} + \frac{1}{2} \log(\sigma(\mathbf{x}_i)^2), \quad (3)$$

where θ denotes the network parameters, f denotes the models and the corresponding output is $\{f(\mathbf{x}_i), \sigma\}$. The aleatoric uncertainty on data element $\mathbf{x}_i$ is denoted by $\sigma(\mathbf{x}_i)^2$, which is roughly the same as σ^2 in the previous section. The deviation originally brought by training data can be learned in an unsupervised manner. ConvNet is regularized with MC dropout to avoid overfitting, which means the network tries to fit a trend instead of remember every single element during training. The deviation $\|\mathbf{y}_i - f(\mathbf{x}_i)\|^2$ can be recognized as the noise originally brought by training data, which is divided by $\sigma(\mathbf{x}_i)^2$ to offset the noise. In other words, the cost function tries to offset the noise through dividing $\|\mathbf{y}_i - f(\mathbf{x}_i)\|^2$ by $\sigma(\mathbf{x}_i)^2$. However, the network tends to predict all σ^2 to be infinitely large to minimize the cost function if we eliminate $\sigma(\mathbf{x}_i)^2$ only. A regularizer $\frac{1}{2} \log(\sigma(\mathbf{x}_i)^2)$ is added to the cost function to avoid this phenomenon. The network learns to output high values of $\sigma(\mathbf{x}_i)^2$ when $\|\mathbf{y}_i - f(\mathbf{x}_i)\|^2$ is large enough in this way. For every single training dataset $\mathbf{x}$, the model outputs a result $f(\mathbf{x}_i)$ and the aleatoric uncertainty $\sigma(\mathbf{x}_i)^2$, which is also used to calculate epistemic uncertainty.

The network does not output aleatoric uncertainty directly in our implementation since the cost function becomes illegal when there is no noise in a single element

($\sigma(\mathbf{x}_i)^2 = 0$). We make the network output as $\log(\sigma^2)$ to avoid this phenomenon and the cost function becomes:

$$L(\theta) = \frac{1}{N} \sum_i \frac{1}{2} \exp(-\log(\sigma^2)) \|\mathbf{y}_i - f(\mathbf{x}_i)\|^2 + \frac{1}{2} \log(\sigma^2), \quad (4)$$

and the two kinds of uncertainty is usually added to characterize the deviation between training and testing set and noise contained within the training set.

A threshold of uncertainty is set according to the expected reliability of segmentation, and results with uncertainty higher than the threshold are dropped to reduce the number of false segmentations. Although it is possible to drop some true positive results in very rare cases, it is still helpful in segmentation accuracy improvement.

3.3 Data augmentation

Motivated by 'mixup' technology [32], data augmentation is used to manually increase the size of the training dataset for medical image segmentation to deal with the issue of inadequate training samples. Specifically, in our implementation, a rotation variant is utilized to create training images that closely resemble one particular feature, e.g., tumors in human organs.

The uncertainty reduction method is explained in detail with training with computed tomography (CT) images. Typical training data for lung nodule analysis are slice-by-slice images captured with a certain spacing. For example, the LUNA16 dataset contains 888 CT images selected from the LiDC/IDRI dataset. Most of them have very similar details since the equipment scans the human body for one slice in a step of 1.3 degrees rotated around the Z-axis.

We combine images from adjacent slices to create mixup samples. A mixup sample is a linear combination of two training features: (x_i, y_i) and (x_j, y_j) , which is given by:

$$\begin{aligned} \mathbf{x}_{mixup} &= \lambda x_i + (1 - \lambda) x_j, \\ \mathbf{y}_{mixup} &= \lambda y_i + (1 - \lambda) y_j, \end{aligned} \quad (5)$$

where the parameter λ lies between $[0, 1]$ and obeys the beta distribution $\beta(\alpha, \alpha)$, where $\alpha \in (0, \infty)$. Instead of choosing x_{ij} at random, we select features from images captured from adjacent slices. According to the data provided in [32], $\alpha \in [0.1, 0.4]$ leads to the best performance in data augmentation of medical image segmentation. A smaller value of α leads to overfitting in training, and the reverse condition results in underfitting.

A typical neural network relies on training backpropagation to update the loss function parameters. Assume the effect of training from the i -th voxel's value is denoted by p_i , and the problem transfers to investigating whether the

equality in Eq. (6) is established. If it is established, it may suggest that mixing samples increase the batch size and, in turn, increase the generalization ability of the learned network.

$$\frac{\partial L}{\partial p_i}(\lambda x_i + (1 - \lambda)x_j) = \lambda \frac{\partial L}{\partial p_i}(x_i) + (1 - \lambda) \frac{\partial L}{\partial p_i}(x_j). \quad (6)$$

We try different combinations until we establish Eq. (6) and train a network that has the highest generalization ability. As a result, we have the best training sample characterizations given the same quantity of labeled training data, which may have the best segmentation performance.

4 Results and discussions

We carry out evaluations to validate uncertainty reduction improvement in terms of voxel-level and lesion-level accuracy by means of the true positive rate (TPR) and false detection rate (FDR). It is a widely accepted way in medical image segmentation since a patient's lesion count are indicative of disease and progression. Specifically, we evaluate our method with two state-of-the-art segmentation uncertainty estimation methods: probabilistic U-Net [21] and segmentation variability estimation [10]. The proposed method is evaluated with a single prediction, which does not contain any uncertainty reduction mechanism, as the baseline. The segmentation accuracy with the proposed method is studied on publicly available benchmarks from potential pneumonia patients. The results are compared with officially provided annotations and categorized into true positive (TP), false positive (FP), true negative (TN) and false negative (FN) results. The results are counted to calculate the true positive rate (TPR) and false detection rate (FDR) indexes as follows:

$$TPR = \frac{TP}{TP + FN}, FDR = 1 - \frac{TP}{TP + FP}, \quad (7)$$

respectively. As a quantitative measurement, we also evaluate the results with the Dice accuracy index [35] as:

$$Dice = \frac{2 \times TP}{2 \times TP + FN + FP}, \quad (8)$$

which is essentially a per-voxel detection score that evaluates the degree of overlap between the segmentation mask and ground truth mask. The higher the Dice score is, the better the performance it indicates. A perfect segmentation result has 100% coverage and yields a Dice score of 1.

Since the Dice index cannot reflect the effect of uncertainty estimation directly, we adopt the metric "uncertainty gap" proposed in [31], which is a measurement of

the variance difference between right and wrong predictions. Generally, the expected result would be low uncertainty values in the right segmented voxels/lesions and high uncertainty in the wrong ones. Therefore, the uncertainty gap is supposed to be positively correlated, which means that the higher the better.

4.1 Data and implementation

We test our method and competitors in two benchmarks: the Lung Nodule Analysis 16 (LUNA16) challenge [36] and the Liver Tumor Segmentation challenge (LiTS) [37]. Originating from the LiDC-IDRI dataset, LUNA16 contains CT images from lung slices smaller than 3 mm. They can be recognized as pulmonary nodules of different rotations, which is ideal for data augmentation in the training of neural networks. Since LUNA16 has 10 equally distributed samples, we select one of them as the test set and the remaining 9 sets as training samples for cross validation. The final result is a combination of 10 results from cross validations. The LiTS challenge contains 130 sets of training samples and 70 sets for testing. It is subtracted from the 3D Image Reconstruction for Comparison of Algorithm Database (3Dircadb), which includes patient 3D medical images and the manual segmentation of the various structures of interest performed by clinical experts. It is famous for the small number of training samples and large variations in data quality, appearance and spacing.

In the training of segmentation network, we adopt U-Net [27] in probabilistic U-Net and our method and FCN [38] in segmentation variability estimation. Adaptive moment estimation (Adam) [39] is used to adjust the learning rate, which is initialized as 10^{-3} with a batch size of 5. In one training epoch with PyTorch, we choose $\alpha = 0.4$ in mixup to increase the generalization ability of the trained models.

4.2 Voxel-level analysis

The segmentation on 3D medical images results in a group of predicted voxels with foreground or background attribute. This voxel-level metric provides an understanding of the neural network on a voxel-by-voxel performance for each scan. Different from pixels in 2D images, voxels in 3D images contain thickness information. For example, the CT images in LUNA16 have a resolution of 512×512 and a voxel depth of 3 mm. The sigmoid output of the network is binarized for segmentation with a threshold, which is varied to explore the range of operation points that the network reaches. Then the voxel-wise counts of TP, FP and FN are accumulated for a given image and calculated for TPR/FDR data. To see if the uncertainty estimations are useful in the medical image segmentation, these metrics are

plotted in a receiver operating characteristic (ROC) curves for the retained voxels at different uncertainty threshold. Voxels with certain level of uncertainty are filtered to calculate TPR/FDR data and a new ROC curve is plotted with the metrics. We try different level of uncertainty in filtering, which is a common approach reported in related academic literature [10], and each threshold is correspond to the percentage of voxels involved in metrics calculation.

The ROC curves for different thresholds of uncertainties at the voxel-level evaluation are illustrated in Fig. 2. The curves suggest that the TPR gradually increases with more and more segmented voxels with low certainty is being dropped, which means uncertainties in predicted voxels are reduced with the augmentation of the training data. The TPR increases gradually when the FDR gradually levels off to 0.5, and an increase in the TP result causes more FP detection. Specifically, the TPR increase is more obvious at the beginning of the test, but the slope of the curve decreases as more FP results are accumulated. Comparing with the baseline curve, a notably high TPR of the ROC curve is attributed to the significant low voxel-level retention. It demonstrates that the network makes incorrect predictions on uncertain voxels, and it is a strong indicator that the uncertainty measures are useful. Since the threshold is selected based on the ratio of retained voxels, a higher threshold is expected to result in lower TPR value. But it is not always true in our implementation because increase of FDR is slower when a more “strict” rule is being used with more and more voxels are being detected. The reason for this phenomenon is that the decrease on FP

detection may be quicker than TP with more and more detection are dropped, which is another indicator that uncertainty estimation is useful in image segmentation. However, we are not intended to set the threshold in a very low level. It needs to be set according to the characteristics of the training and testing samples as well as the expected segmentation reliability, which finds a balance between accuracy and segmentation coverage.

4.3 Lesion-level analysis

As one of the most important applications of medical image segmentation, clinical diagnosis focuses on lesion detection. We are particularly interested in the segmented voxels associated with lesions. The network performance of detection on lesions can be assessed with a lesion-level TPR/FDR. According to the *de facto* standard in clinical practice, ground lesions smaller than 3 voxels are ignored in the annotated ground truth to obtain the lesion-level results from segmentation. As a widely accepted standard used in [10], we recognize a lesion as TP detection when the segmented voxel and its 18-connected neighborhood overlaps at least 50% with the ground truth lesion voxels. A lesion is recognized as FP when 3 or more voxels do not overlap with ground truth and insufficient overlap results in FN result. The calculation of TPR/FDR remains the same as the voxel-level analysis. Currently, TPs, FPs and FNs are associated with large and small lesion detection in an equal weight manner. A range of operating points are filtered to obtain lesion-level ROC curves. Same as the

(a) LUNA16

(b) LiTS

Fig. 2 ROC curves of voxel-level performance on the retained predictions for different uncertainty thresholds. Percentage in the brackets shows ratio of voxels retained at that particular threshold.

old. The uncertainty threshold used to generate a particular color's curve is the same across plots in different dataset

voxel-level analysis, the uncertainty level is calculated for each candidate lesion. The ROC curve reflects a higher TPR and lower FDR when the most uncertain lesions are mainly consist of the FP results.

The ROC curves for different uncertainty thresholds in lesion-level segmentation are illustrated in Fig. 3. The curves suggest that the filtered voxels with low certainty contribute to reduction in FP and FN detection, which in turn improvements in lesion segmentation at all operating points, even if the minimum proportion of uncertainties is reduced. The neural network does not perform well on small lesions, while the uncertainty reduction filters out the most uncertain voxels in segmentation, which are the lesion boundaries. This leads to a small overlapping ratio when compared with the annotated ground truth in the benchmark. The curve in LUNA16 has slightly higher TPR than LiTS since the LiTS dataset contains more small lesions, so removal of falsely segmented lesions improves the overall performance. We can see that the percentage of lesions retained in LiTS is lower than that of LUNA16. The two datasets have fewer reserved detection comparing with voxel-level analysis in the whole because a more strict rule is used. The curves after uncertainty reduction have similar trends when compared with the ROC curve without uncertainty reduction. This phenomenon shows that the lesion-level ROC curve represents the segmentation performance better than the voxel-level curve. The isolated filtered segmented voxels also reduce the number of TP results, which counter reflects the improvement of the TPR value. According to a manual check, the lesion-level TPR improves by 0.5% when only 2% of the most

uncertain results are filtered. ROC curves have higher TPR and lower FDR across the uncertainty reduction areas, which indicate more precise prediction and segmentation.

4.4 Quantitative result

As a quantitative measurement in validating medical volume segmentation, the Dice score is accumulated in all three methods with and without data augmentation. To study the effect of MC simulation iterations on segmentation accuracy, we try different iterations of MC dropout N . We found that the accuracy improves gradually with an increase in N during training and reaches the highest level when $N = 20$.

The quantitative evaluations of segmentation uncertainty from different estimation methods are summarized in Table 2. To make the annotation clear, the competitors are replaced by the corresponding network structures. It can be observed that the data augmentation reduces epistemic uncertainty in models trained with all three methods, which levels up the Dice score by 1-2%

Table 2 Dice scores (%) accumulated with three methods in different network frameworks with and without data augmentation

Dataset	Aug.	FCN [38]	P-UNet [21]	Our method
LUNA16	No	91.1 ± 3.8	90.7 ± 4.3	90.3 ± 4.8
	Yes	92.8 ± 3.1	93.1 ± 2.9	92.9 ± 3.2
LiTS	No	83.1 ± 3.4	84.0 ± 3.9	82.9 ± 3.2
	Yes	85.9 ± 1.1	84.3 ± 3.7	84.5 ± 2.6

(a) LUNA16

(b) LiTS

Fig. 3 ROC curves of lesion-level performance on the retained predictions for different uncertainty thresholds. Percentage in the brackets shows ratio of lesions retained at that particular thresh-

old. The uncertainty threshold used to generate a particular color's curve is the same across plots in different dataset

Table 3 Uncertainty gap accumulated with three methods in different network structures in LUNA16 and LiTS challenge

Dataset	Method	Right Pre.	Wrong Pre.	Gap
LUNA16	FCN	0.95	0.86	0.09
	P-UNet	0.78	0.74	0.04
	Our	0.94	0.83	0.11
LiTS	FCN	0.91	0.79	0.12
	P-UNet	0.87	0.81	0.06
	Our	0.95	0.78	0.17

on average. Since the LUNA16 benchmark ignores lung nodules smaller than 3 mm, the overall performance on the LUNA16 benchmark is better than that on LiTS. One possible explanation is that the trained models are more uncertain on lesion contours, and the small lesion part contains more boundary information, which causes more voxel removal. Consequently, the segmented results have less overlap with the part manually labeled by a radiologist and lower degree of overlap between the segmentation mask and ground truth mask.

The uncertainty gap between correct and incorrect predictions is summarized in Table 3. We can see that our method gains a higher uncertainty gap in two tests with LUNA16 and LiTS compared with segmentation variability estimation and P-UNet. This suggests that the prediction on uncertain areas is closer to the ground truth in our method. The prediction uncertainty values in the LUNA16 challenge are lower than those of LiTS with P-UNet and our method, which means that segmentation with ConvNet has fewer uncertainties than LiTS. One possible reason is that the removal of extremely small lesions makes segmentation easier to implement. The only exception is the uncertainty gap with FCN, where the correct and incorrect predictions are slightly lower in LiTS than in LUNA16. However, the trend in the uncertainty gap remains the same as that of the other two methods. This suggests that the influence of dataset diversity is higher than that of network models.

4.5 Qualitative result

The comparison of segmentation on LiTS challenge are collected in this section to illustrate the result in a more intuitive way. Figure 4 compares several examples of uncertainties reduced segmentation results from several

representative volumes with the corresponding ground truth.

The figure suggests that the segmentation areas are obviously smaller than the ground truth, which means the uncertain voxels are those on which the network makes incorrect predictions on. In general, the lesion contours contain more uncertain voxels where false positive detections are frequently made. This phenomenon does not go beyond our expectations since delineating lesion boundaries is a challenging task, and even radiologist experts may disagree on the precise boundaries of lesions and organs. Images in one row come from the same slice of the same volume, we can see that the segmented areas are smaller than ground truth and the green areas in the ground truth cover more space than the red mask in the middle column. The narrow boundary of human liver covers fewer region in the light red mask than deep red mask. The reason for this phenomenon is that the segment voxels with lower confidence are removed in uncertainty reduction. This removal is crucial in obtaining accurate lesion segmentation. Benefiting from our data augmentation method, the segmented organs match better with the ground truth, especially in small axial image cross sections.

5 Conclusion

Focusing on epistemic and aleatoric uncertainties in training and testing ConvNet models for medical image segmentation, we proposed a data augmentation method with rotation variation from images captured on different slices of organs. The two types of uncertainties are quantized and a proper mechanism is adopted to enrich the training samples, which facilitates learning more texture information with the same quantity of labeled training data. Experiments demonstrate that the proposed method is helpful in reducing the uncertainties produced during data collection and training. It provides the fundamental basis for diagnosis applications with medical images segmentation.

In the future, we will integrate the uncertainty measures obtained from MC dropout samples into a loss function and explore other epistemic uncertainty measurements in medical image segmentation. We will also investigate the task-specific advantages of different measures, which will provide significant advancement in image segmentation on one type of image modality.

Acknowledgements This work was supported by China National Key R&D Program under Grant No. 2017YFC1703506, National Natural Science Foundation of China under Grant No. 81703946, Henan Science and Technology Research under Grant No. 212102310362, Excellent Young Talents Fund Program of Higher Education Institutions of Henan Province under Grant No. 2020GGJS104 and

Fig. 4 Semantic segmentation samples extracted from transverse sections from LiTS challenge by our method. Images in the first and second columns with light red masks indicate segmentation results of liver and tumor, respectively. The segmentation with deep red color in the third column is the ground truth information provided by LiTS dataset. The slice and volume number of capture is denoted in captions of each subfigure. The image and segmentation are displayed by iTK-SNAP [40]

Doctoral Scientific Research Foundation of HACTCM under Grant No. RSBSJJ2020-25.

Declarations

Conflict of interest The authors declare that they have no competing interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Sharma N, Aggarwal LM (2010) Automated medical image segmentation techniques. *J Med Phys* 35(1):3–14
- Prassni JS, Ropinski T, Hinrichs K (2011) Uncertainty-aware guided volume segmentation. *IEEE Trans Vis Comput Graph* 16(6):1358–1365
- Kamnitsas K, Ledig C, Newcombe VF, Simpson JP, Kane AD, Menon DK, Rueckert D, Glocker B (2016) Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Med Image Anal* 36:61
- Gibson E, Giganti F, Hu Y, Bonmati E, Bandula S, Gurusamy K, Davidson B, Pereira SP, Clarkson MJ, Barratt DC (2018) Automatic multi-organ segmentation on abdominal ct with dense v-networks. *IEEE Trans Med Imaging* 37(8):1822–1834
- Yu L, Zhang Z, Li X, Xing L (2021) Deep sinogram completion with image prior for metal artifact reduction in ct images. *IEEE Trans Med Imaging* 40(1):228–238
- Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542(7639):115–118

7. Chen S, Ma K, Zheng Y (2019) Med3d: transfer learning for 3d medical image analysis. [arXiv:abs/1904.00625](https://arxiv.org/abs/1904.00625)
8. Wang G, Li W, Aertsen M, Deprest J, Ourselin S, Vercauteren T (2019) Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing* 338(APR.21):34–45
9. Hüllermeier E, Waegeman W (2021) Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach Learn* 110:457–506
10. Nair T, Precup D, Arnold DL, Arbel T (2020) Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Med Image Anal* 59:655–663
11. Roy AG, Conjeti S, Navab N, Wachinger C (2018) Inherent brain segmentation quality control from fully convnet Monte Carlo sampling. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G (eds) *Proceedings of medical image computing and computer-assisted intervention*. Springer International Publishing, pp 664–672
12. Kendall A, Gal Y (2017) What uncertainties do we need in bayesian deep learning for computer vision? In: *Proceedings of the international conference on neural information processing systems*. Curran Associates Inc, Red Hook, NY, USA, pp 5580–5590
13. Li W, Wang G, Fidon L, Ourselin S, Cardoso MJ, Vercauteren T (2017). On the compactness, efficiency, and representation of 3d convolutional networks: brain parcellation as a pretext task. In: *Proceedings of international conference on information processing in medical imaging*, pp 348–360
14. Zhu Y, Zabaras N (2018) Bayesian deep convolutional encoder-decoder networks for surrogate modeling and uncertainty quantification. *J Comput Phys* 366:415–447
15. Neal RM (1996) *Bayesian learning for neural networks*, volume 118, chapter Lecture notes in statistics. Springer
16. Teye M, Azizpour H, Smith K (2018) Bayesian uncertainty estimation for batch normalized deep networks. In: *Proceedings of medical image computing and computer-assisted intervention*
17. Graves A (2011) Practical variational inference for neural networks. In: *Proceedings of the international conference on neural information processing systems, NIPS'11*. Curran Associates Inc, Red Hook, NY, USA, pp 2348–2356
18. Louizos C, Welling M (2016) Structured and efficient variational deep learning with matrix gaussian posteriors. In: *Proceedings of the international conference on international conference on machine learning*, vol 48, pp 1708–1716
19. Lakshminarayanan B, Pritzel A, Blundell C (2017) Simple and scalable predictive uncertainty estimation using deep ensembles. In: *Proceedings of international conference on neural information processing systems, NIPS*. Curran Associates Inc, Red Hook, NY, USA, pp 6405–6416
20. Rupperecht C, Laina I, DiPietro R, Baust M, Tombari F, Navab N, Hager GD (2017) Learning in an uncertain world: representing ambiguity through multiple hypotheses. In: *Proceedings of the IEEE international conference on computer vision*, pp 3591–3600
21. Kohl SA, Romera-Paredes B, Meyer C, De Fauw J, Ledsam JR, Maier-Hein KH, Eslami SM, Rezende DJ, Ronneberger O (2018) A probabilistic u-net for segmentation of ambiguous images. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, vol 31. pp 6965–6975
22. Kohl SA, Romera-Paredes B, Maier-Hein KH, Rezende DJ, Eslami SM, Kohli P, Zisserman A, Ronneberger O (2019) A hierarchical probabilistic u-net for modeling multi-scale ambiguities. [arXiv:1905.13077](https://arxiv.org/abs/1905.13077)
23. Baumgartner CF, Tezcan KC, Chaitanya K, Hötter AM, Muehlmatter UJ, Schawkat K, Becker AS, Donati O, Konukoglu E (2019) Phiseg: capturing uncertainty in medical image segmentation. In: Shen D, Liu T, Peters TM, Staib LH, Essert C, Zhou S, Yap PT, Khan A (eds) *Medical image computing and computer assisted intervention*. Springer International Publishing, Cham, pp 119–127
24. Pan H, Feng Y, Chen Q, Meyer C, Feng X (2019) Prostate segmentation from 3d mri using a two-stage model and variable-input based uncertainty measure. In: *Proceedings of international symposium on biomedical imaging*, pp 468–471
25. Durasov N, Bagautdinov T, Baque P, Fua P (2021). Masksembles for uncertainty estimation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 13539–13548
26. Osband I, Blundell C, Pritzel A, Van Roy B (2016) Deep exploration via bootstrapped dqn. In: *Proceedings of international conference on neural information processing systems, NIPS*. Curran Associates Inc, Red Hook, NY, USA, pp 4033–4041
27. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O (2016) 3d u-net: learning dense volumetric segmentation from sparse annotation. In: Ourselin S, Joskowicz L, Sabuncu MR, Unal G, Wells W (eds) *Proceedings of medical image computing and computer-assisted intervention*. Springer International Publishing, Cham, pp 424–432
28. Ozdemir O, Woodward B, Berlin A (2017) Propagating uncertainty in multi-stage bayesian convolutional neural networks with application to pulmonary nodule detection. In: *Proceedings of the international conference on neural information processing systems*
29. Bragman FJS, Tanno R, Eaton-Rosen Z, Li W, Hawkes DJ, Ourselin S, Alexander DC, McClelland JR, Cardoso MJ (2018) Uncertainty in multitask learning: joint representations for probabilistic mr-only radiotherapy planning. In: *Proceedings of medical image computing and computer-assisted intervention*, pp 1712–1783
30. Eaton-Rosen Z, Bragman FJS, Bisdas S, Ourselin S, and Cardoso MJ (2018). Towards safe deep learning: accurately quantifying biomarker uncertainty in neural network predictions. In: *Proceedings of medical image computing and computer-assisted intervention*
31. Zheng Z, Yang Y (2011) Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *Int J Comput Vision* 129:1106–1120
32. Zhang H, Cisse M, Dauphin Y, Lopez-Paz D (2018) Mixup: beyond empirical risk minimization. In: *Proceedings of international conference on learning representations*, pp 1–13
33. Gal Y (2016) *Uncertainty in deep learning*. phd thesis, University of Cambridge
34. Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR (2012) Improving neural networks by preventing co-adaptation of feature detectors. *Comput Sci* 3(4):212–223
35. Taha AA, Hanbury A (2015) Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging* 15:29
36. Setio AAA, Traverso A, de Bel Thomas et al (2016) Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Med Image Anal* 42:1–13
37. Bilic P, Christ PF et al (2019) The liver tumor segmentation benchmark (lits). *Computer Research Repository*. [arXiv:abs/1901.04056](https://arxiv.org/abs/1901.04056)
38. Shelhamer E, Long J, Darrell T (2017) Fully convolutional networks for semantic segmentation. *IEEE Trans Pattern Anal Mach Intell* 39(4):640–651
39. Kingma DP, Ba J (2015). Adam: a method for stochastic optimization. In: *Proceedings of international conference on learning representations*, pp 1–15
40. Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, Gerig G (2006) User-guided 3D active contour segmentation

of anatomical structures: significantly improved efficiency and reliability. *Neuroimage* 31(3):1116–1128

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.