

# Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks

Guotai Wang<sup>a,b,c\*</sup>, Wenqi Li<sup>a,b</sup>, Michael Aertsen<sup>d</sup>, Jan Deprest<sup>a,d,e,f</sup>, Sébastien Ourselin<sup>b</sup>, Tom Vercauteren<sup>a,b,f</sup>

<sup>a</sup>Wellcome / EPSRC Centre for Interventional and Surgical Sciences, University College London, London, UK

<sup>b</sup>School of Biomedical Engineering and Imaging Sciences, King's College London, London, UK

<sup>c</sup>School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China

<sup>d</sup>Department of Radiology, University Hospitals Leuven, Leuven, Belgium

<sup>e</sup>Institute for Women's Health, University College London, London, UK

<sup>f</sup>Department of Obstetrics and Gynaecology, University Hospitals Leuven, Leuven, Belgium

## Abstract

Despite the state-of-the-art performance for medical image segmentation, deep convolutional neural networks (CNNs) have rarely provided uncertainty estimations regarding their segmentation outputs, e.g., model (*epistemic*) and image-based (*aleatoric*) uncertainties. In this work, we analyze these different types of uncertainties for CNN-based 2D and 3D medical image segmentation tasks at both pixel level and structure level. We additionally propose a test-time augmentation-based *aleatoric* uncertainty to analyze the effect of different transformations of the input image on the segmentation output. Test-time augmentation has been previously used to improve segmentation accuracy, yet not been formulated in a consistent mathematical framework. Hence, we also propose a theoretical formulation of test-time augmentation, where a distribution of the prediction is estimated by Monte Carlo simulation with prior distributions of parameters in an image acquisition model that involves image transformations and noise. We compare and combine our proposed *aleatoric* uncertainty with model uncertainty. Experiments with segmentation of fetal brains and brain tumors from 2D and 3D Magnetic Resonance Images (MRI) showed that 1) the test-time augmentation-based *aleatoric* uncertainty provides a better uncertainty estimation than calculating the test-time dropout-based model uncertainty alone and helps to reduce overconfident incorrect predictions, and 2) our test-time augmentation outperforms a single-prediction baseline and dropout-based multiple predictions.

**Keywords:** Uncertainty estimation, convolutional neural networks, medical image segmentation, data augmentation

## 1. Introduction

Segmentation of medical images is an essential task for many applications such as anatomical structure modeling, tumor growth measurement, surgical planing and treatment assessment (Sharma and Aggarwal, 2010). Despite the breadth and depth of current research, it is very challenging to achieve accurate and reliable segmentation results for many targets (Withey and Koles, 2007). This is often due to poor image quality, inhomogeneous appearances brought by pathology, various imaging protocols and large variations of the segmentation target among patients. Therefore, uncertainty estimation of segmentation results is critical for understanding how reliable the segmentations are. For example, for many images, the segmentation results of pixels near the boundary are likely to be uncertain because of the low contrast between the segmentation target and surrounding tissues, where uncertainty information of the segmentation can be used to indicate potential mis-segmented regions or guide user interactions for refinement (Prassni et al., 2010; Wang et al., 2018b).

In recent years, deep learning with convolutional neural networks (CNN) has achieved the state-of-the-art performance for many medical image segmentation tasks (Milletari et al., 2016; Abdulkadir et al., 2016; Kamnitsas et al., 2017). Despite their impressive performance and the ability of automatic feature learning, these approaches do not by default provide uncertainty estimation for their segmentation results. In addition, having access to a large training set plays an important role for deep CNNs to achieve human-level performance (Esteva et al., 2017; Rajpurkar et al., 2017). However, for medical image segmentation tasks, collecting a very large dataset with pixel-wise annotations for training is usually difficult and time-consuming. As a result, current medical image segmentation methods based on deep CNNs use relatively small datasets compared with those for natural image recognition (Russakovsky et al., 2015). This is likely to introduce more uncertain predictions for the segmentation results, and also leads to uncertainty of downstream analysis, such as volumetric measurement of the target. Therefore, uncertainty estimation is highly desired for deep CNN-based medical image segmentation methods.

Several works have investigated uncertainty estimation for deep neural networks (Kendall and Gal, 2017; Lakshminarayanan et al., 2017; Zhu and Zabarar, 2018; Ayhan and Berens, 2018).

\*Corresponding author

Email address: guotai.1.wang@kcl.ac.uk (Guotai Wang<sup>a,b,c</sup>)

They focused mainly on image classification or regression tasks, where the prediction outputs are high-level image labels or bounding box parameters, therefore uncertainty estimation is usually only given for the high-level predictions. In contrast, pixel-wise predictions are involved in segmentation tasks, therefore pixel-wise uncertainty estimation is highly desirable. In addition, in most interactive segmentation cases, pixel-wise uncertainty information is more helpful for intelligently guiding the user to give interactions. However, previous works have rarely demonstrated uncertainty estimation for deep CNN-based medical image segmentation. As suggested by Kendall and Gal (2017), there are two major types of predictive uncertainties for deep CNNs: *epistemic* uncertainty and *aleatoric* uncertainty. *Epistemic* uncertainty is also known as model uncertainty that can be explained away given enough training data, while *aleatoric* uncertainty depends on noise or randomness in the input testing image.

In contrast to previous works focusing mainly on classification or regression-related uncertainty estimation, and recent works of Nair et al. (2018) and Roy et al. (2018) investigating only test-time dropout-based (*epistemic*) uncertainty for segmentation, we extensively investigate different kinds of uncertainties for CNN-based medical image segmentation, including not only *epistemic* but also *aleatoric* uncertainties for this task. We also propose a more general estimation of *aleatoric* uncertainty that is related to not only image noise but also spatial transformations of the input, considering different possible poses of the object during image acquisition. To obtain the transformation-related uncertainty, we augment the input image at test time, and obtain an estimation of the distribution of the prediction based on test-time augmentation. Test-time augmentation (e.g., rotation, scaling, flipping) has been recently used to improve performance of image classification (Matsunaga et al., 2017) and nodule detection (Jin et al., 2018). Ayhan and Berens (2018) also showed its utility for uncertainty estimation in a fundus image classification task. However, the previous works have not provided a mathematical or theoretical formulation for this. Motivated by these observations, we propose a mathematical formulation for test-time augmentation, and analyze its performance for the general *aleatoric* uncertainty estimation in medical image segmentation tasks. In the proposed formulation, we represent an image as a result of an acquisition process which involves geometric transformations and image noise. We model the hidden parameters of the image acquisition process with prior distributions, and predict the distribution of the output segmentation for a test image with a Monte Carlo sampling process. With the samples from the distribution of the predictive output based on the same pre-trained CNN, the variance/entropy can be calculated for these samples, which provides an estimation of the *aleatoric* uncertainty for the segmentation.

The contribution of this work is three-fold. First, we propose a theoretical formulation of test-time augmentation for deep learning. Test-time augmentation has not been mathematically formulated by existing works, and our proposed mathematical formulation is general for image recognition tasks. Second, with the proposed formulation of test-time augmentation,

we propose a general *aleatoric* uncertainty estimation for medical image segmentation, where the uncertainty comes from not only image noise but also spatial transformations. Third, we analyze different types of uncertainty estimation for the deep CNN-based segmentation, and validate the superiority of the proposed general *aleatoric* uncertainty with both 2D and 3D segmentation tasks.

## 2. Related Works

### 2.1. Segmentation Uncertainty

Uncertainty estimation has been widely investigated for many existing medical image segmentation tasks. As way of examples, Saad et al. (2010) used shape and appearance prior information to estimate the uncertainty for probabilistic segmentation of medical imaging. Shi et al. (2011) estimated the uncertainty of graph cut-based cardiac image segmentation, which was used to improve the robustness of the system. Sankaran et al. (2015) estimated lumen segmentation uncertainty for realistic patient-specific blood flow modeling. Parisot et al. (2014) used segmentation uncertainty to guide content-driven adaptive sampling for concurrent brain tumor segmentation and registration. Prassni et al. (2010) visualized the uncertainty of a random walker-based segmentation to guide volume segmentation of brain Magnetic Resonance Images (MRI) and abdominal Computed Tomography (CT) images. Top et al. (2011) combined uncertainty estimation with active learning to reduce user time for interactive 3D image segmentation.

### 2.2. Uncertainty Estimation for Deep CNNs

For deep CNNs, both *epistemic* and *aleatoric* uncertainties have been investigated in recent years. For model (*epistemic*) uncertainty, exact Bayesian networks offer a mathematically grounded method, but they are hard to implement and computationally expensive. Alternatively, it has been shown that dropout at test time can be cast as a Bayesian approximation to represent model uncertainty (Gal and Ghahramani, 2016; Li et al., 2017). Zhu and Zabaras (2018) used Stochastic Variational Gradient Descent (SVGD) to perform approximate Bayesian inference on uncertain CNN parameters. A variety of other approximation methods such as Markov chain Monte Carlo (MCMC) (Neal, 2012), Monte Carlo Batch Normalization (MCBN) (Teye et al., 2018) and variational Bayesian methods (Graves, 2011; Louizos and Welling, 2016) have also been developed. Lakshminarayanan et al. (2017) proposed ensembles of multiple models for uncertainty estimation, which was simple and scalable to implement. For test image-based (*aleatoric*) uncertainty, Kendall and Gal (2017) proposed a unified Bayesian deep learning framework to learn mappings from input data to *aleatoric* uncertainty and composed them with *epistemic* uncertainty, where the *aleatoric* uncertainty was modeled as learned loss attenuation and further categorized into *homoscedastic* uncertainty and *heteroscedastic* uncertainty. Ayhan and Berens (2018) used test-time augmentation for *aleatoric* uncertainty estimation, which was an efficient and effective way to explore the locality of a testing sample. However, its utility for medical image segmentation has not been demonstrated.

### 2.3. Test-Time Augmentation

Data augmentation was originally proposed for the training of deep neural networks. It was employed to enlarge a relatively small dataset by applying transformations to its samples to create new ones for training (Krizhevsky et al., 2012). The transformations for augmentation typically include flipping, cropping, rotating, and scaling training images. Abdulkadir et al. (2016) and Ronneberger et al. (2015) also used elastic deformations for biomedical image segmentation. Several studies have empirically found that combining predictions of multiple transformed versions of a test image helps to improve the performance. For example, Matsunaga et al. (2017) geometrically transformed test images for skin lesion classification. Radosavovic et al. (2017) used a single model to predict multiple transformed copies of unlabeled images for data distillation. Jin et al. (2018) tested on samples extended by rotation and translation for pulmonary nodule detection. However, all these works used test-time augmentation as an ad hoc method, without detailed formulation or theoretical explanation, and did not apply it to uncertainty estimation for segmentation tasks.

## 3. Method

The proposed general *aleatoric* uncertainty estimation is formulated in a consistent mathematical framework including two parts. The first part is a mathematical representation of ensembles of predictions of multiple transformed versions of the input. We represent an image as a result of an image acquisition model with hidden parameters in Section 3.1. Then we formulate test-time augmentation as inference with hidden parameters following given prior distributions in Section 3.2. The second part calculates the diversity of the prediction results of an augmented test image, and it is used to estimate the *aleatoric* uncertainty related to image transformations and noise. This is detailed in Section 3.3. Our proposed *aleatoric* uncertainty is compared and combined with *epistemic* uncertainty, which is described in Section 3.4. Finally, we apply our proposed method to structure-wise uncertainty estimation in Section 3.5.

### 3.1. Image Acquisition Model

The image acquisition model describes the process by which the observed images have been obtained. This process is confronted with a lot of factors that can be related or unrelated to the imaged object, such as blurring, down-sampling, spatial transformation, and system noise. While blurring and down-sampling are commonly considered for image super-resolution (Yue et al., 2016), in the context of image recognition they have a relatively lower impact. Therefore, we focus on the spatial transformation and noise, and highlight that adding more complex intensity changes or other forms of data augmentation such as elastic deformations is a straightforward extension. The image acquisition model is:

$$X = \mathcal{T}_\beta(X_0) + \mathbf{e} \quad (1)$$

where  $X_0$  is an underlying image in a certain position and orientation, i.e., a hidden variable.  $\mathcal{T}$  is a transformation operator

that is applied to  $X_0$ .  $\beta$  is the set of parameters of the transformation, and  $\mathbf{e}$  represents the noise that is added to the transformed image.  $X$  denotes the observed image that is used for inference at test time. Though the transformations can be in spatial, intensity or feature space, in this work we only study the impact of reversible spatial transformations (e.g., flipping, scaling, rotation and translation), which are the most common types of transformations occurring during image acquisition and used for data augmentation purposes. Let  $\mathcal{T}_\beta^{-1}$  denote the inverse transformation of  $\mathcal{T}_\beta$ , then we have:

$$X_0 = \mathcal{T}_\beta^{-1}(X - \mathbf{e}) \quad (2)$$

Similarly to data augmentation at training time, we assume that the distribution of  $X$  covers the distribution of  $X_0$ . In a given application, this assumption leads to some prior distributions of the transformation parameters and noise. For example, in a 2D slice of fetal brain MRI, the orientation of the fetal brain can span all possible directions in a 2D plane, therefore the rotation angle  $\mathbf{r}$  can be modeled with a uniform prior distribution  $\mathbf{r} \sim U(0, 2\pi)$ . The image noise is commonly modeled as a Gaussian distribution, i.e.,  $\mathbf{e} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ , where  $\boldsymbol{\mu}$  and  $\boldsymbol{\sigma}$  are the mean and standard deviation respectively. Let  $p(\beta)$  and  $p(\mathbf{e})$  represent the prior distribution of  $\beta$  and  $\mathbf{e}$  respectively, therefore we have  $\beta \sim p(\beta)$  and  $\mathbf{e} \sim p(\mathbf{e})$ .

Let  $Y$  and  $Y_0$  be the labels related to  $X$  and  $X_0$  respectively. For image classification,  $Y$  and  $Y_0$  are categorical variables, and they should be invariant with regard to transformations and noise, therefore  $Y = Y_0$ . For image segmentation,  $Y$  and  $Y_0$  are discretized label maps, and they are equivariant with the spatial transformation, i.e.,  $Y = \mathcal{T}_\beta(Y_0)$ .

### 3.2. Inference with Hidden Variables

In the context of deep learning, let  $f(\cdot)$  be the function represented by a neural network, and  $\boldsymbol{\theta}$  represent the parameters learned from a set of training images with their corresponding annotations. In a standard formulation, the prediction  $Y$  of a test image  $X$  is inferred by:

$$Y = f(\boldsymbol{\theta}, X) \quad (3)$$

For regression problems,  $Y$  refers to continuous values. For segmentation or classification problems,  $Y$  refers to discretized labels obtained by *argmax* operation in the last layer of the network. Since  $X$  is only one of many possible observations of the underlying image  $X_0$ , direct inference with  $X$  may lead to a biased result affected by the specific transformation and noise associated with  $X$ . To address this problem, we aim at inferring it with the help of the latent  $X_0$  instead:

$$Y = \mathcal{T}_\beta(Y_0) = \mathcal{T}_\beta(f(\boldsymbol{\theta}, X_0)) = \mathcal{T}_\beta(f(\boldsymbol{\theta}, \mathcal{T}_\beta^{-1}(X - \mathbf{e}))) \quad (4)$$

where the exact values of  $\beta$  and  $\mathbf{e}$  for  $X$  are unknown. Instead of finding a deterministic prediction of  $X$ , we alternatively consider the distribution of  $Y$  for a robust inference given the dis-

tributions of  $\beta$  and  $e$ .

$$p(Y|X) = p\left(\mathcal{T}_\beta\left(f(\theta, \mathcal{T}_\beta^{-1}(X - e))\right)\right), \text{ where } \beta \sim p(\beta), e \sim p(e) \quad (5)$$

For regression problems, we obtain the final prediction for  $X$  by calculating the expectation of  $Y$  using the distribution  $p(Y|X)$ .

$$\begin{aligned} E(Y|X) &= \int y p(y|X) dy \\ &= \int_{\beta \sim p(\beta), e \sim p(e)} \mathcal{T}_\beta\left(f(\theta, \mathcal{T}_\beta^{-1}(X - e))\right) p(\beta) p(e) d\beta de \end{aligned} \quad (6)$$

Calculating  $E(Y|X)$  with Eq. (6) is computationally expensive, as  $\beta$  and  $e$  may take continuous values and  $p(\beta)$  is a complex joint distribution of different types of transformations. Alternatively, we estimate  $E(Y|X)$  by using Monte Carlo simulation. Let  $N$  represent the total number of simulation runs. In the  $n$ -th simulation run, the prediction is:

$$y_n = \mathcal{T}_{\beta_n}\left(f(\theta, \mathcal{T}_{\beta_n}^{-1}(X - e_n))\right) \quad (7)$$

where  $\beta_n \sim p(\beta)$ ,  $e_n \sim p(e)$ . To obtain  $y_n$ , we first randomly sample  $\beta_n$  and  $e_n$  from the prior distributions  $p(\beta)$  and  $p(e)$ , respectively. Then we obtain one possible hidden image with  $\beta_n$  and  $e_n$  based on Eq. (2), and feed it into the trained network to get its prediction, which is transformed with  $\beta_n$  to obtain  $y_n$  according to Eq. (4). With the set  $\mathcal{Y} = \{y_1, y_2, \dots, y_N\}$  sampled from  $p(Y|X)$ ,  $E(Y|X)$  is estimated as the average of  $\mathcal{Y}$  and we use it as the final prediction  $\hat{Y}$  for  $X$ :

$$\hat{Y} = E(Y|X) \approx \frac{1}{N} \sum_{n=1}^N y_n \quad (8)$$

For classification or segmentation problems,  $p(Y|X)$  is a discretized distribution. We obtain the final prediction for  $X$  by maximum likelihood estimation:

$$\hat{Y} = \arg \max_y p(y|X) \approx \text{Mode}(\mathcal{Y}) \quad (9)$$

where  $\text{Mode}(\mathcal{Y})$  is the most frequent element in  $\mathcal{Y}$ . This corresponds to majority voting of multiple predictions.

### 3.3. Aleatoric Uncertainty Estimation with Test-Time Augmentation

The uncertainty is estimated by measuring how diverse the predictions for a given image are. Both the variance and entropy of the distribution  $p(Y|X)$  can be used to estimate uncertainty. However, variance is not sufficiently representative in the context of multi-modal distributions. In this paper we use entropy for uncertainty estimation:

$$H(Y|X) = - \int p(y|X) \ln(p(y|X)) dy \quad (10)$$

With the Monte Carlo simulation in Section 3.2, we can approximate  $H(Y|X)$  from the simulation results  $\mathcal{Y} = \{y_1, y_2, \dots, y_N\}$ .

Suppose there are  $M$  unique values in  $\mathcal{Y}$ . For classification tasks, this typically refers to  $M$  labels. Assume the frequency of the  $m$ -th unique value is  $\hat{p}_m$ , then  $H(Y|X)$  is approximated as:

$$H(Y|X) \approx - \sum_{m=1}^M \hat{p}_m \ln(\hat{p}_m) \quad (11)$$

For segmentation tasks, pixel-wise uncertainty estimation is desirable. Let  $Y^i$  denote the predicted label for the  $i$ -th pixel. With the Monte Carlo simulation, a set of values for  $Y^i$  are obtained  $\mathcal{Y}^i = \{y_1^i, y_2^i, \dots, y_N^i\}$ . The entropy of the distribution of  $Y^i$  is therefore approximated as:

$$H(Y^i|X) \approx - \sum_{m=1}^M \hat{p}_m^i \ln(\hat{p}_m^i) \quad (12)$$

where  $\hat{p}_m^i$  is the frequency of the  $m$ -th unique value in  $\mathcal{Y}^i$ .

### 3.4. Epistemic Uncertainty Estimation

To obtain model (*epistemic*) uncertainty estimation, we follow the test-time dropout method proposed by Gal and Ghahramani (2016). In this method, let  $q(\theta)$  be an approximating distribution over the set of network parameters  $\theta$  with its elements randomly set to zero according to Bernoulli random variables.  $q(\theta)$  can be achieved by minimizing the Kullback-Leibler divergence between  $q(\theta)$  and the posterior distribution of  $\theta$  given a training set. After training, the predictive distribution of a test image  $X$  can be expressed as:

$$p(Y|X) = \int p(Y|X, \omega) q(\omega) d\omega \quad (13)$$

The distribution of the prediction can be sampled based on Monte Carlo samples of the trained network (i.e, MC dropout):  $y_n = f(\theta_n, X)$  where  $\theta_n$  is a Monte Carlo sample from  $q(\theta)$ . Assume the number of samples is  $N$ , and the sampled set of the distribution of  $Y$  is  $\mathcal{Y} = \{y_1, y_2, \dots, y_N\}$ . The final prediction for  $X$  can be estimated by Eq. (8) for regression problems or Eq. (9) for classification/segmentation problems. The *epistemic* uncertainty estimation can therefore be calculated based on variance or entropy of the sampled  $N$  predictions. To keep consistent with our *aleatoric* uncertainty, we use entropy for this purpose, which is similar to Eq. (12). Test-time dropout may be interpreted as a way of ensembles of networks for testing. In the work of Lakshminarayanan et al. (2017), ensembles of neural networks was explicitly proposed as an alternative solution of test-time dropout for estimating *epistemic* uncertainty.

### 3.5. Structure-wise Uncertainty Estimation

Nair et al. (2018) and Roy et al. (2018) used Monte Carlo samples generated by test-time dropout for structure/lesion-wise uncertainty estimation. Following these works, we extend the structure-wise uncertainty estimation method by using Monte Carlo samples generated by not only test-time dropout, but also test-time augmentation described in 3.2. For  $N$  samples from the Monte Carlo simulation, let  $\mathcal{V} = \{v_1, v_2, \dots, v_N\}$  denote the

set of volumes of the segmented structure, where  $v_i$  is the volume of the segmented structure in the  $i$ -th simulation. Let  $\mu_{\mathcal{V}}$  and  $\sigma_{\mathcal{V}}$  denote the mean value and standard deviation of  $\mathcal{V}$  respectively. We use the volume variation coefficient (VVC) to estimate the structure-wise uncertainty:

$$VVC = \frac{\sigma_{\mathcal{V}}}{\mu_{\mathcal{V}}} \quad (14)$$

where VVC is agnostic to the size of the segmented structure.

## 4. Experiments and Results

We validated our proposed testing and uncertainty estimation method with two segmentation tasks: 2D fetal brain segmentation from MRI slices and 3D brain tumor segmentation from multi-modal MRI volumes. The implementation details for 2D and 3D segmentation are described in Section 4.1 and Section 4.2 respectively.

In both tasks, we compared different types of uncertainties for the segmentation results: 1) the proposed *aleatoric* uncertainty based on our formulated test-time augmentation (TTA), 2) the *epistemic* uncertainty based on test-time dropout (TTD) described in Section 3.4, and 3) hybrid uncertainty that combines the *aleatoric* and *epistemic* uncertainties based on TTA + TTD. For each of these three methods, the uncertainty was obtained by Eq. (12) with  $N$  predictions. For TTD and TTA + TTD, the dropout probability was set as a typical value of 0.5 (Gal and Ghahramani, 2016).

We also evaluated the segmentation accuracy of these different prediction methods: TTA, TTD, TTA + TTD and the baseline that uses a single prediction without TTA and TTD. For a given training set, all these methods used the same model that was trained with data augmentation and dropout at training time. The augmentation during training followed the same formulation in Section 3.1. We investigated the relationship between each type of uncertainty and segmentation error in order to know which uncertainty has a better ability to indicate potential mis-segmentations. Quantitative evaluations of segmentation accuracy are based on Dice score and Average Symmetric Surface Distance (ASSD).

$$Dice = \frac{2 \times TP}{2 \times TP + FN + FP} \quad (15)$$

where  $TP$ ,  $FP$  and  $FN$  are true positive, false positive and false negative respectively. The definition of ASSD is:

$$ASSD = \frac{1}{|S| + |G|} \left( \sum_{s \in S} d(s, G) + \sum_{g \in G} d(g, S) \right) \quad (16)$$

where  $S$  and  $G$  denote the set of surface points of a segmentation result and the ground truth respectively.  $d(s, G)$  is the shortest Euclidean distance between a point  $s \in S$  and all the points in  $G$ .

### 4.1. 2D Fetal Brain Segmentation from MRI

Fetal MRI has been increasingly used for study of the developing fetus as it provides a better soft tissue contrast than the

widely used prenatal sonography. The most commonly used imaging protocol for fetal MRI is Single-Shot Fast Spin Echo (SSFSE) that acquires images in a fast speed and mitigates the effect of fetal motion, leading to stacks of thick 2D slices. Segmentation is a fundamental step for fetal brain study, e.g., it plays an important role in inter-slice motion correction and high-resolution volume reconstruction (Tourbier et al., 2017; Ebner et al., 2018). Recently, CNNs have achieved the state-of-the-art performance for 2D fetal brain segmentation (Rajchl et al., 2016; Salehi et al., 2017, 2018). In this experiment, we segment the 2D fetal brain using deep CNNs with uncertainty estimation.

#### 4.1.1. Data and Implementation

We collected clinical T2-weighted MRI scans of 60 fetuses in the second trimester with SSFSE on a 1.5 Tesla MR system (Aera, Siemens, Erlangen, Germany). The data for each fetus contained three stacks of 2D slices acquired in axial, sagittal and coronal views respectively, with pixel size 0.63 mm to 1.58 mm and slice thickness 3 mm to 6 mm. The gestational age ranged from 19 weeks to 33 weeks. We used 2640 slices from 120 stacks of 40 patients for training, 278 slices from 12 stacks of 4 patients for validation and 1180 slices from 48 stacks of 16 patients for testing. Two radiologists manually segmented the brain region for all the stacks slice-by-slice, where one radiologist gave a segmentation first, and then the second senior radiologist refined the segmentation if disagreement existed, the output of which were used as the ground truth. We used this dataset for two reasons. First, our dataset fits with a typical medical image segmentation application where the number of annotated images is limited. This leads the uncertainty information to be of high interest for robust prediction and our downstream tasks such as fetal brain reconstruction and volume measurement. Second, the position and orientation of fetal brain have large variations, which is suitable for investigating the effect of data augmentation. For preprocessing, we normalized each stack by its intensity mean and standard deviation, and resampled each slice with pixel size 1.0 mm.

We used 2D networks of Fully Convolutional Network (FCN) (Long et al., 2015), U-Net (Ronneberger et al., 2015) and P-Net (Wang et al., 2018b). The networks were implemented in TensorFlow<sup>1</sup> (Abadi et al., 2016) using NiftyNet<sup>2</sup> (Li et al., 2017; Gibson et al., 2018). During training, we used Adaptive Moment Estimation (Adam) to adjust the learning rate that was initialized as  $10^{-3}$ , with batch size 5, weight decay  $10^{-7}$  and iteration number  $10k$ . We represented the transformation parameter  $\beta$  in the proposed augmentation framework as a combination of  $f_i$ ,  $r$  and  $s$ , where  $f_i$  is a random variable for flipping along each 2D axis,  $r$  is the rotation angle in 2D, and  $s$  is a scaling factor. The prior distributions of these transformation parameters and random intensity noise were modeled as  $f_i \sim \text{Bern}(\mu_f)$ ,  $r \sim U(r_0, r_1)$ ,  $s \sim U(s_0, s_1)$  and  $e \sim N(\mu_e, \sigma_e)$ . The hyper-parameters for our fetal brain segmentation task were set as

<sup>1</sup><https://www.tensorflow.org>

<sup>2</sup><http://www.niftynet.io>

$\mu_f = 0.5$ ,  $r_0 = 0$ ,  $r_1 = 2\pi$ ,  $s_0 = 0.8$  and  $s_1 = 1.2$ . For the random noise, we set  $\mu_e = 0$  and  $\sigma_e = 0.05$ , as a median-filter smoothed version of a normalized image in our dataset has a standard deviation around 0.95. We augmented the training data with this formulation, and during test time, TTA used the same prior distributions of augmentation parameters as used for training.

#### 4.1.2. Segmentation Results with Uncertainty

Fig. 1 shows a visual comparison of different types of uncertainties for segmentation of three fetal brain images in coronal, sagittal and axial view respectively. The results were based on the same trained model of U-Net with train-time augmentation, and the Monte Carlo simulation number  $N$  was 20 for TTD, TTA, and TTA + TTD to obtain *epistemic*, *aleatoric* and hybrid uncertainties respectively. In each subfigure, the first row presents the input and the segmentation obtained by the single-prediction baseline. The other rows show these three types of uncertainties and their corresponding segmentation results respectively. The uncertainty maps in odd columns are represented by pixel-wise entropy of  $N$  predictions and encoded by the color bar in the left top corner. In the uncertainty maps, purple pixels have low uncertainty values and yellow pixels have high uncertainty values. Fig. 1(a) shows a fetal brain in coronal view. In this case, the baseline prediction method achieved a good segmentation result. It can be observed that for *epistemic* uncertainty calculated by TTD, most of the uncertain segmentations are located near the border of the segmented foreground, while the pixels with a larger distance to the border have a very high confidence (i.e., low uncertainty). In addition, the *epistemic* uncertainty map contains some random noise in the brain region. In contrast, the *aleatoric* uncertainty obtained by TTA contains less random noise and it shows uncertain segmentations not only on the border but also in some challenging areas in the lower right corner, as highlighted by the white arrows. In that region, the result obtained by TTA has an over-segmentation, and this is corresponding to the high values in the same region of the *aleatoric* uncertainty map. The hybrid uncertainty calculated by TTA + TTD is a mixture of *epistemic* and *aleatoric* uncertainty. As shown in the last row of Fig. 1(a), it looks similar to the *aleatoric* uncertainty map except for some random noise.

Fig. 1(b) and Fig. 1(c) show two other cases where the single-prediction baseline obtained an over-segmentation and an under-segmentation respectively. It can be observed that the *epistemic* uncertainty map shows a high confidence (low uncertainty) in these mis-segmented regions. This leads to a lot of overconfident incorrect segmentations, as highlighted by the white arrows in Fig. 1(b) and (c). In comparison, the *aleatoric* uncertainty map obtained by TTA shows a larger uncertain area that is mainly corresponding to mis-segmented regions of the baseline. In these two cases, The hybrid uncertainty also looks similar to the *aleatoric* uncertainty map. The comparison indicates that the *aleatoric* uncertainty has a better ability than the *epistemic* uncertainty to indicate mis-segmentations of non-border pixels. For these pixels, the segmentation output is more affected by different transformations of the input (*aleatoric*) rather than

variations of model parameters (*epistemic*).

Fig. 1(b) and (c) also show that TTD using different model parameters seemed to obtain very little improvement from the baseline. In comparison, TTA using different input transformations corrected the large mis-segmentations and achieved a more noticeable improvement from the baseline. It can also be observed that the results obtained by TTA + TTD are very similar to those obtained by TTA, which shows TTA is more suitable to improving the segmentation than TTD.

#### 4.1.3. Quantitative Evaluation

To quantitatively evaluate the segmentation results, we measured Dice score and ASSD of predictions by different testing methods with three network structures: FCN (Long et al., 2015), U-Net (Ronneberger et al., 2015) and P-Net (Wang et al., 2018b). For all of these CNNs, we used data augmentation at training time to enlarge the training set. At inference time, we compared the baseline testing method (without Monte Carlo simulation) with TTD, TTA and TTA + TTD. We first investigated how the segmentation accuracy changes with the increase of the number of Monte Carlo simulation runs  $N$ . The results measured with all the testing images are shown in Fig. 2. We found that for all of these three networks, the segmentation accuracy of TTD remains close to that of the single-prediction baseline. For TTA and TTA + TTD, an improvement of segmentation accuracy can be observed when  $N$  increases from 1 to 10. When  $N$  is larger than 20, the segmentation accuracy for these two methods reaches a plateau.

In addition to the previous scenario using augmentation at both training and test time, we also evaluated the performance of TTD and TTA when data augmentation was not used for training. The quantitative evaluations of combinations of different training methods and testing methods ( $N=20$ ) are shown in Table 1. It can be observed that for both training with and without data augmentation, TTA has a better ability to improve the segmentation accuracy than TTD. Combining TTA and TTD can further improve the segmentation accuracy, but it does not significantly outperform TTA ( $p$ -value  $> 0.05$ ).

Fig. 3 shows Dice distributions of five example stacks of fetal brain MRI. The results were based on the same trained model of U-Net with train-time augmentation. Note that the baseline had only one prediction for each image, and the Monte Carlo simulation number  $N$  was 20 for TTD, TTA and TTA + TTD. It can be observed that for each case, the Dice of TTD is distributed closely around that of the baseline. In comparison, the Dice distribution of TTA has a higher average than that of TTD, indicating TTA's better ability of improving segmentation accuracy. The results of TTA also have a larger variance than that of TTD, which shows TTA can provide more structure-wise uncertainty information. Fig. 3 also shows that the performance of TTA + TTD is close to that of TTA.

#### 4.1.4. Correlation between Uncertainty and Segmentation Error

To investigate how our uncertainty estimation methods can indicate incorrect segmentation, we measured the uncertainty and segmentation error at both pixel-level and structure-level.

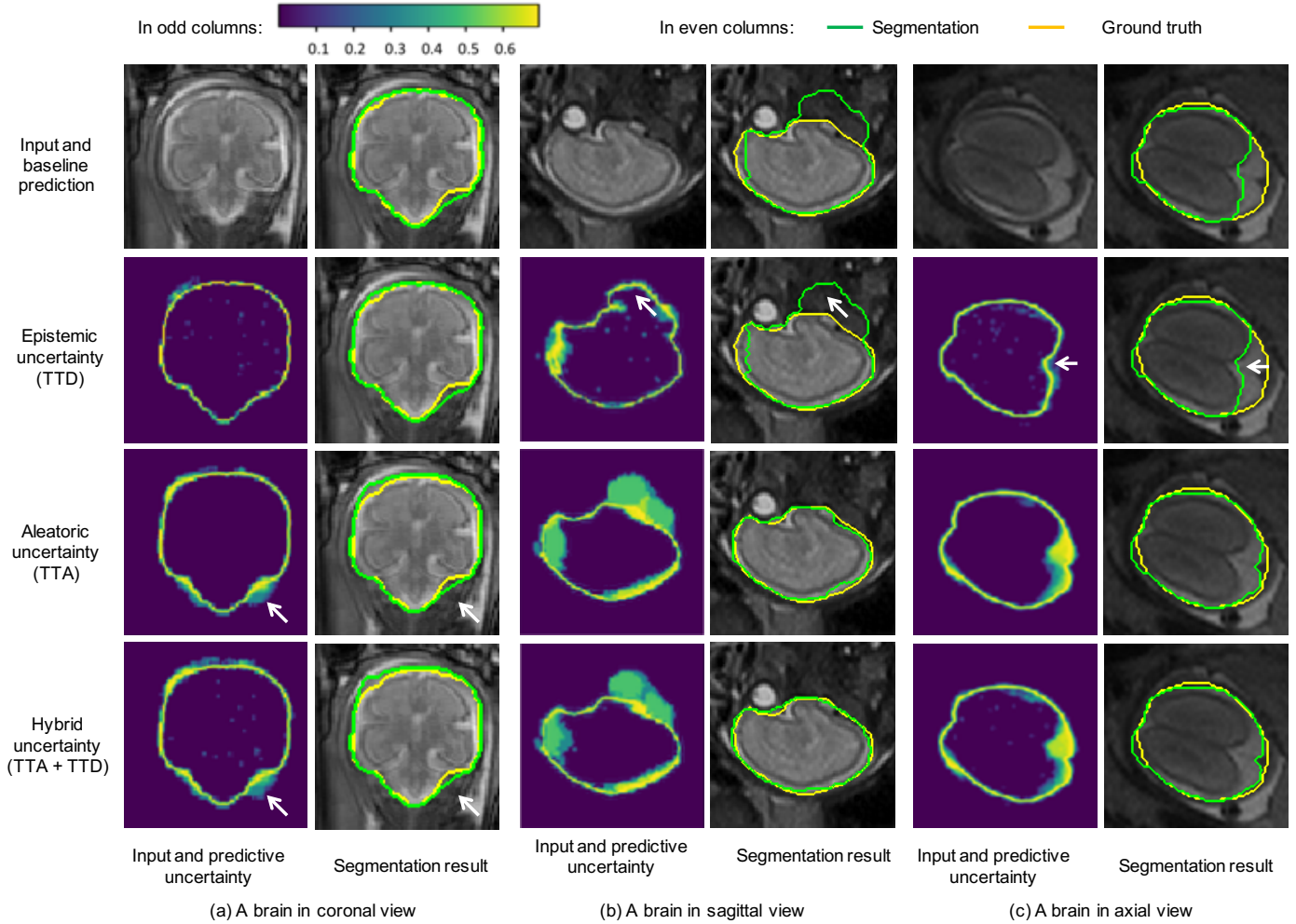


Figure 1: Visual comparison of different types of uncertainties and their corresponding segmentations for fetal brain. The uncertainty maps in odd columns are based on Monte Carlo simulation with  $N = 20$  and encoded by the color bar in the left up corner (low uncertainty shown in purple and high uncertainty shown in yellow). The white arrows in (a) show the *aleatoric* and hybrid uncertainties in a challenging area, and the white arrows in (b) and (c) show mis-segmented regions with very low *epistemic* uncertainty. TTD: test-time dropout, TTA: test-time augmentation.

For pixel-level evaluation, we measured the joint histogram of pixel-wise uncertainty and error rate for TTD, TTA, and TTA + TTD respectively. The histogram was obtained by statistically calculating the error rate of pixels at different pixel-wise uncertainty levels in each slice. The results based on U-Net with  $N = 20$  are shown in Fig. 4, where the joint histograms have been normalized by the number of total pixels in the testing images for visualization. For each type of pixel-wise uncertainty, we calculated the average error rate at each pixel-wise uncertainty level, leading to a curve of error rate as a function of pixel-wise uncertainty, i.e., the red curves in Fig. 4. This figure shows that the majority of pixels have a low uncertainty with a small error rate. When the uncertainty increases, the error rate also becomes higher gradually. Fig. 4(a) shows the TTD-based uncertainty (*epistemic*). It can be observed that when the prediction uncertainty is low, the result has a steep increase of error rate. In contrast, for the TTA-based uncertainty (*aleatoric*), the increase of error rate is slower, shown in Fig. 4(b). This demonstrates that TTA has fewer overconfident incorrect predictions

than TTD. The dashed ellipses in Fig. 4 also show the different levels of overconfident incorrect predictions for different testing methods.

For structure-wise evaluation, we used VVC to represent structure-wise uncertainty and  $1 - \text{Dice}$  to represent structure-wise segmentation error. Fig. 5 shows the joint distribution of VVC and  $1 - \text{Dice}$  for different testing methods using U-Net trained with data augmentation and  $N = 20$  for inference. The results of TTD, TTA, and TTA + TTD are shown in Fig. 5(a), (b) and (c) respectively. It can be observed that for all the three testing methods, the VVC value tends to become larger when  $1 - \text{Dice}$  grows. However, the slope in Fig. 5(a) is smaller than those in Fig. 5(b) and Fig. 5(c). The comparison shows that TTA-based structure-wise uncertainty estimation is highly related to segmentation error, and TTA leads to a larger scale of VVC than TTD. Combining TTA and TTD leads to similar results to that of TTA.

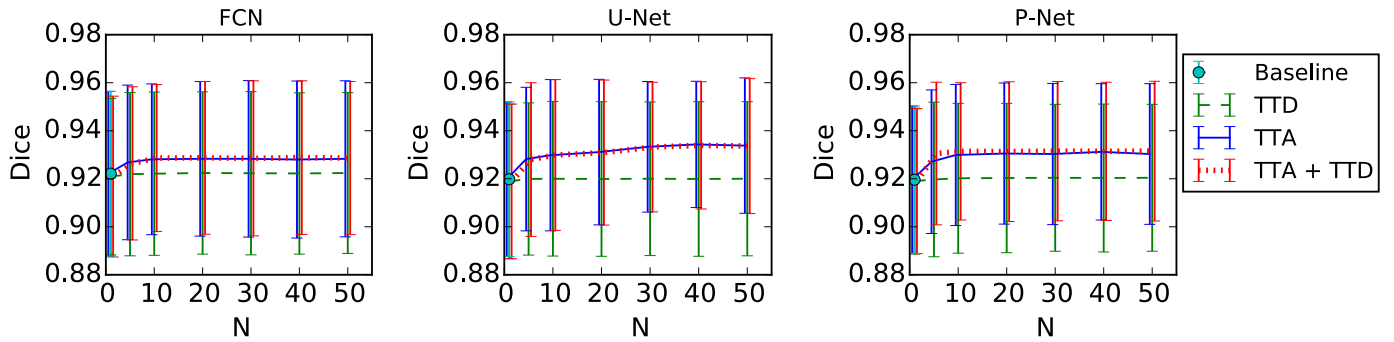


Figure 2: Dice of 2D fetal brain segmentation with different  $N$  that is the number of Monte Carlo simulation runs.

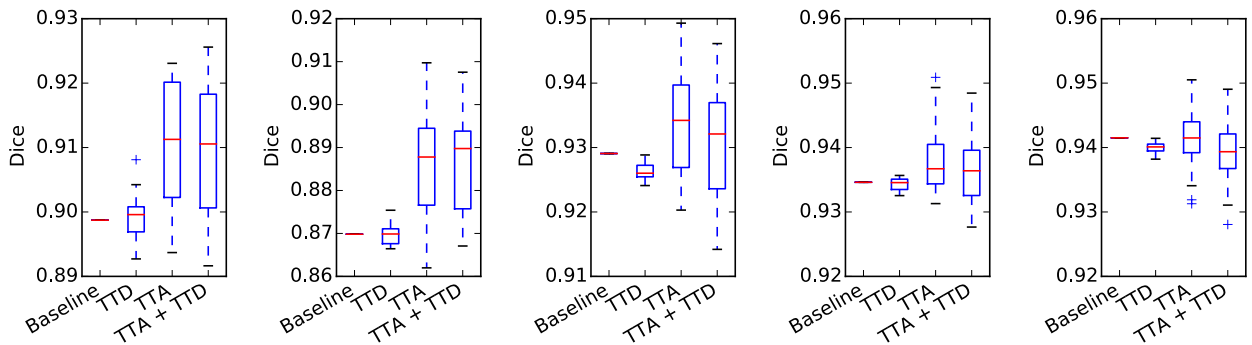


Figure 3: Dice distributions of segmentation results with different testing methods for five example stacks of 2D slices of fetal brain MRI. Note TTA’s higher mean value and variance compared with TTD.

#### 4.2. 3D Brain Tumor Segmentation from Multi-Modal MRI

MRI has become the most commonly used imaging methods for brain tumors. Different MR sequences such as T1-weighted (T1w), contrast enhanced T1-weighted (T1wce), T2-weighted (T2w) and Fluid Attenuation Inversion Recovery (FLAIR) images can provide complementary information for analyzing multiple subregions of brain tumors. Automatic brain tumor segmentation from multi-modal MRI has a potential for better diagnosis, surgical planning and treatment assessment (Menze et al., 2015). Deep neural networks have achieved the state-of-the-art performance on this task (Kamnitsas et al., 2017; Wang et al., 2018a). In this experiment, we analyze the uncertainty of deep CNN-based brain tumor segmentation and show the effect of our proposed test-time augmentation.

##### 4.2.1. Data and Implementation

We used the BraTS 2017<sup>3</sup> (Bakas et al., 2017) training dataset that consisted of volumetric images from 285 studies, with ground truth provided by the organizers. We randomly selected 20 studies for validation and 50 studies for testing, and used the remaining for training. For each study, there were four scans of T1w, T1wce, T2w and FLAIR images, and they had been co-registered. All the images were skull-stripped and re-sampled to an isotropic 1 mm<sup>3</sup> resolution. As a first demonstration of

uncertainty estimation for deep learning-based brain tumor segmentation, we investigate segmentation of the whole tumor from these multi-modal images (Fig. 6). We used 3D U-Net (Abdulkadir et al., 2016), V-Net (Milletari et al., 2016) and W-Net (Wang et al., 2018a) implemented with NiftyNet (Gibson et al., 2018), and employed Adam during training with initial learning rate  $10^{-3}$ , batch size 2, weight decay  $10^{-7}$  and iteration number  $20k$ . W-Net is a 2.5D network, and we compared using W-Net only in axial view and a fusion of axial, sagittal and coronal views. These two implementations are referred to as W-Net(A) and W-Net(ASC) respectively. The transformation parameter  $\beta$  in the proposed augmentation framework consisted of  $f_i$ ,  $r$ ,  $s$  and  $e$ , where  $f_i$  is a random variable for flipping along each 3D axis,  $r$  is the rotation angle along each 3D axis,  $s$  is a scaling factor and  $e$  is intensity noise. The prior distributions were:  $f_i \sim \text{Bern}(0.5)$ ,  $r \sim U(0, 2\pi)$ ,  $s \sim U(0.8, 1.2)$  and  $e \sim N(0, 0.05)$  according to the reduced standard deviation of a median-filtered version of a normalized image. We used this formulated augmentation during training, and also employed it to obtain TTA-based results at test time.

##### 4.2.2. Segmentation Results with Uncertainty

Fig. 6 demonstrates three examples of uncertainty estimation of brain tumor segmentation by different testing methods. The results were based on the same trained model of 3D U-Net (Abdulkadir et al., 2016). The Monte Carlo simulation number  $N$  was 40 for TTD, TTA, and TTA + TTD to obtain

<sup>3</sup><http://www.med.upenn.edu/sbia/brats2017.html>



Table 1: Dice (%) and ASSD (mm) evaluation of 2D fetal brain segmentation with different training and testing methods. Tr-Aug: Training without data augmentation. Tr+Aug: Training with data augmentation. \* denotes significant improvement from the baseline of single prediction in Tr-Aug and Tr+Aug respectively ( $p$ -value  $< 0.05$ ). † denotes significant improvement from Tr-Aug with TTA + TTD ( $p$ -value  $< 0.05$ ).

Train	Test	Dice (%)			ASSD (mm)		
		FCN	U-Net	P-Net	FCN	U-Net	P-Net
Tr-Aug	Baseline	91.05±3.82	90.26±4.77	90.65±4.29	2.68±2.93	3.11±3.34	2.83±3.07
	TTD	91.13±3.60	90.38±4.30	90.93±4.04	2.61±2.85	3.04±2.29	2.69±2.90
	TTA	91.99±3.48*	91.64±4.11*	92.02±3.85*	2.26±2.56*	2.51±3.23*	2.28±2.61*
	TTA + TTD	<b>92.05±3.58*</b>	<b>91.88±3.61*</b>	<b>92.17±3.68*</b>	<b>2.19±2.67*</b>	<b>2.40±2.71*</b>	<b>2.13±2.42*</b>
Tr+Aug	Baseline	92.03±3.44	91.93±3.21	91.98±3.92	2.21±2.52	2.12±2.23	2.32±2.71
	TTD	92.08±3.41	92.00±3.22	92.01±3.89	2.17±2.52	2.03±2.13	2.15±2.58
	TTA	92.79±3.34*	92.88±3.15*	93.05±2.96*	1.88±2.08	1.70±1.75	1.62±1.77*
	TTA + TTD	<b>92.85±3.15*†</b>	<b>92.90±3.16*†</b>	<b>93.14±2.93*†</b>	<b>1.84±1.92</b>	<b>1.67±1.76*†</b>	<b>1.48±1.63*†</b>

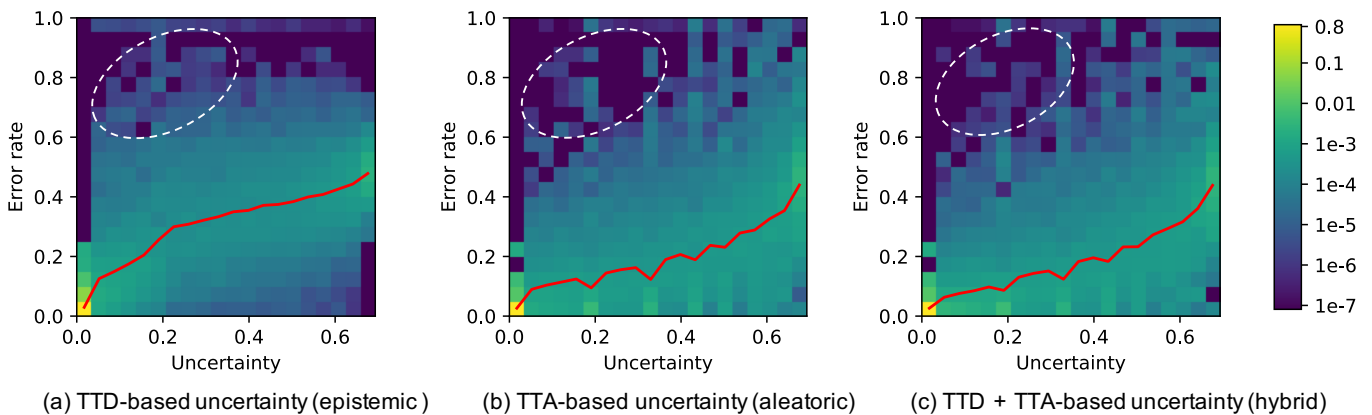


Figure 4: Normalized joint histogram of prediction uncertainty and error rate for 2D fetal brain segmentation. The average error rates at different uncertainty levels are depicted by the red curves. The dashed ellipses show that TTA leads to a lower occurrence of overconfident incorrect predictions than TTD.

*epistemic*, *aleatoric* and hybrid uncertainties respectively. Fig. 6(a) shows a case of high grade glioma (HGG). The baseline of single prediction obtained an over-segmentation at the upper part of the image. The *epistemic* uncertainty obtained by TTD highlights some uncertain predictions at the border of the segmentation and a small part of the over-segmented region. In contrast, the *aleatoric* uncertainty obtained by TTA better highlights the whole over-segmented region, and the hybrid uncertainty map obtained by TTA + TTD is similar to the *aleatoric* uncertainty map. The second column of Fig. 6(a) shows the corresponding segmentations of these uncertainties. It can be observed that the TTD-based result looks similar to the baseline, while TTA and TTA + TTD based results achieve a larger improvement from the baseline. Fig. 6(b) demonstrates another case of HGG brain tumor, and it shows that the over-segmented region in the baseline prediction is better highlighted by TTA-based *aleatoric* uncertainty than TTD-based *epistemic* uncertainty. Fig. 6(c) shows a case of low grade glioma (LGG). The baseline of single prediction obtained an under-segmentation in the middle part of the tumor. The *epistemic* uncertainty obtained by TTD only highlights pixels on the border of the prediction, with a low uncertainty (high confidence) for the under-segmented region. In contrast, the *aleatoric* uncertainty obtained by TTA has a bet-

ter ability to indicate the under-segmentation. The results also show that TTA outperforms TTD for better segmentation.

#### 4.2.3. Quantitative Evaluation

For quantitative evaluations, we calculated the Dice score and ASSD for the segmentation results obtained by the different testing methods that were combined with 3D U-Net (Abdulkadir et al., 2016), V-Net (Milletari et al., 2016) and W-Net (Wang et al., 2018a) respectively. We also compared TTD and TTA with and without train-time data augmentation, respectively. We found that for these networks, the performance of the multi-prediction testing methods reaches a plateau when  $N$  is larger than 40. Table 2 shows the evaluation results with  $N = 40$ . It can be observed that for each network and each training method, multi-prediction methods lead to better performance than the baseline with a single prediction, and TTA outperforms TTD with higher Dice scores and lower ASSD values. Combining TTA and TTD has a slight improvement from using TTA, but the improvement is not significant ( $p$ -value  $< 0.05$ ).

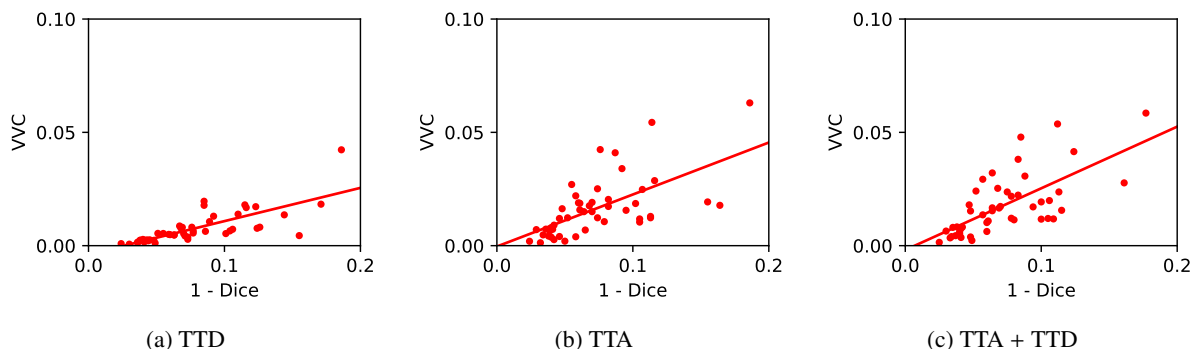


Figure 5: Structure-wise uncertainty in terms of volume variation coefficient (VVC) vs 1-Dice for different testing methods in 2D fetal brain segmentation.

Table 2: Dice (%) and ASSD (mm) evaluation of 3D brain tumor segmentation with different training and testing methods. Tr-Aug: Training without data augmentation. Tr+Aug: Training with data augmentation. W-Net is a 2.5D network and W-Net (ASC) denotes the fusion of axial, sagittal and coronal views according to Wang et al. (2018a). \* denotes significant improvement from the baseline of single prediction in Tr-Aug and Tr+Aug respectively ( $p$ -value < 0.05). † denotes significant improvement from Tr-Aug with TTA + TTD ( $p$ -value < 0.05).

Train	Test	Dice (%)			ASSD (mm)		
		WNet (ASC)	3D U-Net	V-Net	WNet (ASC)	3D U-Net	V-Net
Tr-Aug	Baseline	87.81±7.27	87.26±7.73	86.84±8.38	2.04±1.27	2.62±1.48	2.86±1.79
	TTD	88.14±7.02	87.55±7.33	87.13±8.14	1.95±1.20	2.55±1.41	2.82±1.75
	TTA	89.16±6.48*	88.58±6.50*	87.86±6.97*	1.42±0.93*	1.79±1.16*	1.97±1.40*
	TTA + TTD	<b>89.43±6.14*</b>	<b>88.75±6.34*</b>	<b>88.03±6.56*</b>	<b>1.37±0.89*</b>	<b>1.72±1.23*</b>	<b>1.95±1.31*</b>
Tr+Aug	Baseline	88.76±5.76	88.43±6.67	87.44±7.84	1.61±1.12	1.82±1.17	2.07±1.46
	TTD	88.92±5.73	88.52±6.66	87.56±7.78	1.57±1.06	1.76±1.14	1.99±1.33
	TTA	90.07±5.69*	89.41±6.05*	88.38±6.74*	1.13±0.54*	1.45±0.81	1.67±0.98*
	TTA + TTD	<b>90.35±5.64*†</b>	<b>89.60±5.95*†</b>	<b>88.57±6.32*†</b>	<b>1.10±0.49*</b>	<b>1.39±0.76*†</b>	<b>1.62±0.95*†</b>

#### 4.2.4. Correlation between Uncertainty and Segmentation Error

To study the relationship between prediction uncertainty and segmentation error at voxel-level, we measured voxel-wise uncertainty and voxel-wise error rate at different uncertainty levels. For each of TTD-based (*epistemic*), TTA-based (*aleatoric*) and TTA + TTD-based (hybrid) voxel-wise uncertainty, we obtained the normalized joint histogram of voxel-wise uncertainty and voxel-wise error rate. Fig. 7 shows the results based on 3D U-Net trained with data augmentation and using  $N = 40$  for inference. The red curve shows the average voxel-wise error rate as a function of voxel-wise uncertainty. In Fig. 7(a), the average prediction error rate has a slight change when the TTD-based *epistemic* uncertainty is larger than 0.2. In contrast, Fig. 7(b) and (c) show that the average prediction error rate has a smoother increase with the growth of *aleatoric* and hybrid uncertainties. The comparison demonstrates that the TTA-based *aleatoric* uncertainty leads to fewer over-confident mis-segmentations than the TTD-based *epistemic* uncertainty.

For structure-level evaluation, we also studied the relationship between structure-level uncertainty represented by VVC and structure-level error represented by 1-Dice. Fig. 8 shows their joint distributions with three different testing methods using 3D U-Net. The network was trained with data augmentation, and  $N$  was set as 40 for inference. Fig. 8 shows that TTA-based VVC increases when 1-Dice grows, and the slope

is larger than that of TTD-based VVC. The results of TTA and TTA + TTD are similar, as shown in Fig. 8(b) and (c). The comparison shows that TTA-based structure-wise uncertainty can better indicate segmentation error than TTD-based structure-wise uncertainty.

## 5. Discussion and Conclusion

In our experiments, the number of training images was relatively small compared with many datasets of natural images such as PASCAL VOC, COCO and ImageNet. For medical images, it is typically very difficult to collect a very large dataset for segmentation, as pixel-wise annotations are not only time-consuming to collect but also require expertise of radiologists. Therefore, for most existing medical image segmentation datasets, such as those in Grand challenge<sup>4</sup>, the image numbers are also quite small. Therefore, investigating the segmentation performance of CNNs with limited training data is of high interest for medical image computing community. In addition, our dataset is not very large so that it is suitable for data augmentation, which fits well with our motivation of using data augmentation at training and test time. The need for uncertainty estimation is also stronger in cases where datasets are smaller.

<sup>4</sup> <https://grand-challenge.org/challenges>

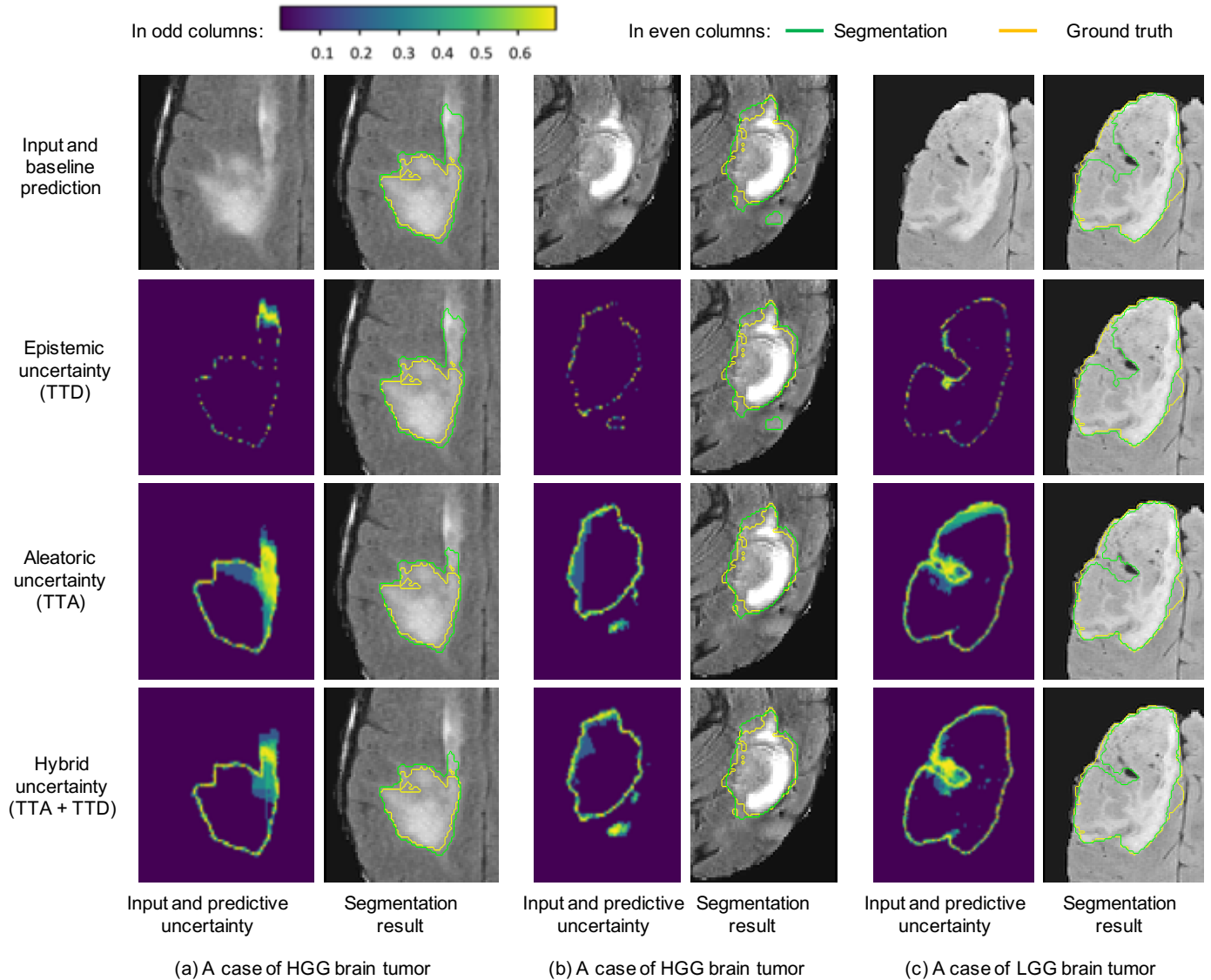


Figure 6: Visual comparison of different testing methods for 3D brain tumor segmentation. The uncertainty maps in odd columns are based on Monte Carlo simulation with  $N = 40$  and encoded by the color bar in the left up corner (low uncertainty shown in purple and high uncertainty shown in yellow). TTD: test-time dropout, TTA: test-time augmentation.

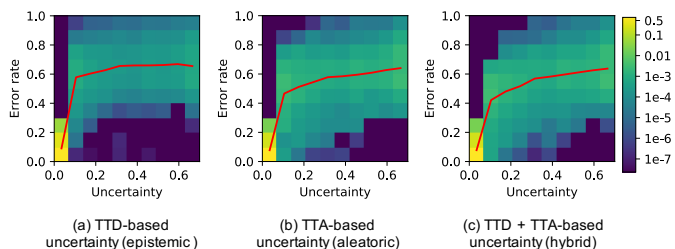


Figure 7: Normalized joint histogram of prediction uncertainty and error rate for 3D brain tumor segmentation. The average error rates at different uncertainty levels are depicted by the red curves.

In our mathematical formulation of test-time augmentation based on an image acquisition model, we explicitly modeled spatial transformations and image noise. However, it can be

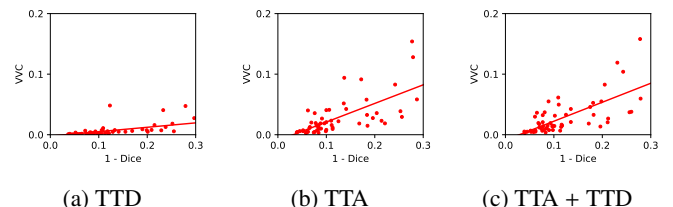


Figure 8: Structure-wise uncertainty in terms of volume variation coefficient (VVC) vs  $1 - \text{Dice}$  for different testing methods in 3D brain tumor segmentation.

easily extended to include more general transformations such as elastic deformations (Abdulkadir et al., 2016) or add a simulated bias field for MRI. In addition to the variation of possible values of model parameters, the prediction result is also dependent on the input data, e.g., image noise and transformations related to the object. Therefore, a good uncertainty estimation should take these factors into consideration. Fig. 1 and 6 show that model uncertainty alone is likely to obtain overconfident incorrect predictions, and TTA plays an important role in reducing such predictions. In Fig. 3 we show five example cases, where each subfigure shows the results for one patient. Table 1 shows the statistical results based on all the testing images. We found that for few testing images TTA +TTD failed to obtain higher Dice scores than TTA, but for the overall testing images, the average Dice of TTA + TTD is slightly larger than that of TTA. Therefore, this leads to the conclusion that TTA + TTD does not always perform better than TTA, and the average performance of TTA + TTD is close to that of TTA, which is also demonstrated in Fig. 1 and 6.

We have demonstrated TTA based on the image acquisition model for image segmentation tasks, but it is general for different image recognition tasks, such as image classification, object detection, and regression. For regression tasks where the outputs are not discretized category labels, the variation of the output distribution might be more suitable than entropy for uncertainty estimation. Table 2 shows the superiority of test-time augmentation for better segmentation accuracy, and it also demonstrates the combination of W-Net in different views helps to improve the performance. This is an ensemble of three networks, and such an ensemble may be used as an alternative for *epistemic* uncertainty estimation, as demonstrated by Lakshminarayanan et al. (2017).

We found that for our tested CNNs and applications, the proper value of Monte Carlo sample  $N$  that leads the segmentation accuracy to a plateau was around 20 to 40. Using an empirical value  $N = 40$  is large enough for our datasets. However, the optimal setting of hyper-parameter  $N$  may change for different datasets. Fixing  $N = 40$  for new applications where the optimal value of  $N$  is smaller would lead to unnecessary computation and reduce efficiency. In some applications where the object has more spatial variations, the optimal  $N$  value may be larger than 40. Therefore, in a new application, we suggest that the optimal  $N$  should be determined by the performance plateau on the validation set.

In conclusion, we analyzed different types of uncertainties for CNN-based medical image segmentation by comparing and combining model (*epistemic*) and input-based (*aleatoric*) uncertainties. We formulated a test-time augmentation-based *aleatoric* uncertainty estimation for medical images that considers the effect of both image noise and spatial transformations. We also proposed a theoretical and mathematical formulation of test-time augmentation, where we obtain a distribution of the prediction by using Monte Carlo simulation and modeling prior distributions of parameters in an image acquisition model. Experiments with 2D and 3D medical image segmentation tasks showed that uncertainty estimation with our formulated TTA helps to reduce overconfident incorrect predictions encountered

by model-based uncertainty estimation and TTA leads to higher segmentation accuracy than a single-prediction baseline and multiple predictions using test-time dropout.

## 6. Acknowledgements

This work was supported by the Wellcome/EPSRC Centre for Medical Engineering [WT 203148/Z/16/Z], an Innovative Engineering for Health award by the Wellcome Trust (WT101957); Engineering and Physical Sciences Research Council (EPSRC) (NS/A000027/1, EP/H046410/1, EP/J020990/1, EP/K005278), Wellcome/EPSRC [203145Z/16/Z], the National Institute for Health Research University College London Hospitals Biomedical Research Centre (NIHR BRC UCLH/UCL), the Royal Society [RG160569], and hardware donated by NVIDIA.

## References

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X., Brain, G., 2016. TensorFlow: A system for large-scale machine learning, in: USENIX Symposium on Operating Systems Design and Implementation, pp. 265–284.
- Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: Learning dense volumetric segmentation from sparse annotation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 424–432.
- Ayhan, M.S., Berens, P., 2018. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks, in: Medical Imaging with Deep Learning, pp. 1–9.
- Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C., 2017. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Nature Scientific Data*, 170117.
- Ebner, M., Wang, G., Li, W., Aertsen, M., Patel, P.A., Aghwane, R., Melbourne, A., Doel, T., David, A.L., Deprest, J., Ourselin, S., Vercauteren, T., 2018. An automated localization, segmentation and reconstruction framework for fetal brain MRI, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 313–320.
- Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 115–118.
- Gal, Y., Ghahramani, Z., 2016. Dropout as a Bayesian approximation: representing model uncertainty in deep learning, in: International Conference on Machine Learning, pp. 1050–1059.
- Gibson, E., Li, W., Sudre, C., Fidon, L., Shaker, D.I., Wang, G., Eaton-Rosen, Z., Gray, R., Doel, T., Hu, Y., Whyntie, T., Nachev, P., Modat, M., Barratt, D.C., Ourselin, S., Cardoso, M.J., Vercauteren, T., 2018. NiftyNet: A deep-learning platform for medical imaging. *Computer Methods and Programs in Biomedicine* 158, 113–122.
- Graves, A., 2011. Practical variational inference for neural networks, in: Advances in Neural Information Processing Systems, pp. 1–9.
- Jin, H., Li, Z., Tong, R., Lin, L., 2018. A deep 3D residual CNN for false positive reduction in pulmonary nodule detection. *Medical Physics* 45, 2097–2107.
- Kamnitsas, K., Ledig, C., Newcombe, V.F.J., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis* 36, 61–78.
- Kendall, A., Gal, Y., 2017. What uncertainties do we need in Bayesian deep learning for computer vision?, in: Advances in Neural Information Processing Systems, pp. 5580–5590.
- Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, pp. 1097–1105.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles, in: *Advances in Neural Information Processing Systems*, pp. 6405–6416.

Li, W., Wang, G., Fidon, L., Ourselin, S., Cardoso, M.J., Vercauteren, T., 2017. On the compactness, efficiency, and representation of 3D convolutional networks: brain parcellation as a pretext task, in: *International Conference on Information Processing in Medical Imaging*, pp. 348–360.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440.

Louizos, C., Welling, M., 2016. Structured and efficient variational deep learning with matrix gaussian posteriors, in: *International Conference on Machine Learning*, pp. 1708–1716.

Matsunaga, K., Hamada, A., Minagawa, A., Koga, H., 2017. Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble. *arXiv preprint arXiv:1703.03108*.

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.A., Arbel, T., Avants, B.B., Ayache, N., Buendia, P., Collins, D.L., Cordier, N., Corso, J.J., Criminisi, A., Das, T., Delingette, H., Demiralp, Ç., Durst, C.R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharruddin, K.M., Jena, R., John, N.M., Konukoglu, E., Lashkari, D., Mariz, J.A., Meier, R., Pereira, S., Precup, D., Price, S.J., Raviv, T.R., Reza, S.M., Ryan, M., Sarikaya, D., Schwartz, L., Shin, H.C., Shotton, J., Silva, C.A., Sousa, N., Subbanna, N.K., Szekely, G., Taylor, T.J., Thomas, O.M., Tustison, N.J., Unal, G., Vasseur, F., Wintermark, M., Ye, D.H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M., Van Leemput, K., 2015. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging* 34, 1993–2024.

Milletari, F., Navab, N., Ahmadi, S.A., 2016. V-Net: Fully convolutional neural networks for volumetric medical image segmentation, in: *International Conference on 3D Vision*, pp. 565–571.

Nair, T., Precup, D., Arnold, D.L., Arbel, T., 2018. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 655–663.

Neal, R.M., 2012. *Bayesian learning for neural networks*. Springer Science & Business Media.

Pariset, S., Wells, W., Chemouny, S., Duffau, H., Paragios, N., 2014. Concurrent tumor segmentation and registration with uncertainty-based sparse non-uniform graphs. *Medical Image Analysis* 18, 647–659.

Prassni, J.S., Ropinski, T., Hinrichs, K., 2010. Uncertainty-aware guided volume segmentation. *IEEE Transactions on Visualization and Computer Graphics* 16, 1358–1365.

Radosavovic, I., Dollár, P., Girshick, R., Gkioxari, G., He, K., 2017. Data distillation: towards omni-supervised learning. *arXiv preprint arXiv:1712.04440*.

Rajchl, M., Lee, M.C.H., Schrans, F., Davidson, A., Passerat-Palmbach, J., Tarroni, G., Alansary, A., Oktay, O., Kainz, B., Rueckert, D., 2016. Learning under distributed weak supervision. *arXiv preprint arXiv:1606.01100*.

Rajpurkar, P., Irvin, J., Bagul, A., Ding, D., Duan, T., Mehta, H., Yang, B., Zhu, K., Laird, D., Ball, R.L., Langlotz, C., Shpanskaya, K., Lungren, M.P., Ng, A., 2017. MURA dataset: towards radiologist-level abnormality detection in musculoskeletal radiographs. *arXiv preprint arXiv:1712.06957*.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241.

Roy, A.G., Conjeti, S., Navab, N., Wachinger, C., 2018. Inherent brain segmentation quality control from fully convnet Monte Carlo sampling, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 664–672.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision* 115, 211–252.

Saad, A., Hamarneh, G., Möller, T., 2010. Exploration and visualization of segmentation uncertainty using shape and appearance prior information. *IEEE Transactions on Visualization and Computer Graphics* 16, 1366–1375.

Salehi, S.S.M., Erdogmus, D., Gholipour, A., 2017. Auto-context convolutional neural network (auto-net) for brain extraction in magnetic resonance imaging. *IEEE Transactions on Medical Imaging* 36, 2319 – 2330.

Salehi, S.S.M., Hashemi, S.R., Velasco-Annis, C., Ouaalam, A., Estroff, J.A., Erdogmus, D., Warfield, S.K., Gholipour, A., 2018. Real-time automatic fetal brain extraction in fetal MRI by deep learning, in: *IEEE International Symposium on Biomedical Imaging*, pp. 720–724.

Sankaran, S., Grady, L., Taylor, C.A., 2015. Fast computation of hemodynamic sensitivity to lumen segmentation uncertainty. *IEEE Transactions on Medical Imaging* 34, 2562–2571.

Sharma, N., Aggarwal, L.M., 2010. Automated medical image segmentation techniques. *Journal of Medical Physics* 35, 3–14.

Shi, W., Zhuang, X., Wolz, R., Simon, D., Tung, K., Wang, H., Ourselin, S., Edwards, P., Razavi, R., Rueckert, D., 2011. A multi-image graph cut approach for cardiac image segmentation and uncertainty estimation, in: *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer Berlin Heidelberg, volume 7085, pp. 178–187.

Teye, M., Azizpour, H., Smith, K., 2018. Bayesian uncertainty estimation for batch normalized deep networks. *arXiv preprint arXiv:1802.06455*.

Top, A., Hamarneh, G., Abugharbich, R., 2011. Active learning for interactive 3D image segmentation, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 603–610.

Tourbier, S., Velasco-Annis, C., Taimouri, V., Haggmann, P., Meuli, R., Warfield, S.K., Bach Cuadra, M., Gholipour, A., 2017. Automated template-based brain localization and extraction for fetal brain MRI reconstruction. *NeuroImage* 155, 460–472.

Wang, G., Li, W., Ourselin, S., Vercauteren, T., 2018a. Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks, in: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Springer International Publishing, pp. 178–190.

Wang, G., Li, W., Zuluaga, M.A., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ourselin, S., Vercauteren, T., 2018b. Interactive medical image segmentation using deep learning with image-specific fine-tuning. *IEEE Transactions on Medical Imaging* 37, 1562–1573.

Withey, D., Koles, Z., 2007. Medical image segmentation: methods and software, in: *Noninvasive Functional Source Imaging of the Brain and Heart and the International Conference on Functional Biomedical Imaging*, pp. 140–143.

Yue, L., Shen, H., Li, J., Yuan, Q., Zhang, H., Zhang, L., 2016. Image super-resolution: the techniques, applications, and future. *Signal Processing* 128, 389–408.

Zhu, Y., Zabarab, N., 2018. Bayesian deep convolutional encoder-decoder networks for surrogate modeling and uncertainty quantification. *arXiv preprint arXiv:1801.06879*.