[Re]Rethinking Bayesian Deep Learning for Semi-Supervised Volumetric Medical Image Segmentation [4]

Project Report by Rohan Banerjee
Mila - Quebec AI Institute and Polytechnique Montreal
rohan.banerjee@polymtl.ca

Abstract

Semi-supervised learning is essential in medical imaging settings where obtaining labeled data for training supervised segmentation models can be difficult. To address this challenge, the authors introduce a novel architecture called generative Bayesian deep learning (GBDL) in paper [4]. GBDL is designed specifically for semi-supervised medical image segmentation and aims to estimate the joint distribution of input medical volumes and their corresponding labels. It consists of three key components: a deep neural network (DNN), a conditional variational autoencoder (cVAE), and a Bayesian inference module, which work together to improve segmentation accuracy using both labeled and unlabeled data.

The proposed GBDL architecture for semi-supervised medical image segmentation offers several advantages over traditional supervised methods. It effectively utilizes both labeled and unlabeled data in the early stages of training, mitigating the risk of overfitting. It also provides uncertainty estimates for each prediction, which is crucial in medical applications. Additionally, GBDL can generate realistic samples from the learned distribution. The paper presents a comprehensive ablation study that validates the effectiveness of each component in GBDL and demonstrates its superior performance compared to state-of-the-art methods on three publicly available datasets.

In addition to the main findings and results, we provide our own interpretation of the paper. This includes our understanding of the proposed method, our efforts in reproducing the reported results, and our own experiments designed to assess the model's ability to generalize in out-of-distribution settings.

1. Introduction

Medical image segmentation is a critical task in medical image analysis, as it plays a key role in various clinical applications such as disease diagnosis, treatment planning, and monitoring. However, unlike natural image segmentation, medical image segmentation often faces the challenge of limited availability of ground truth segmentations for training. This scarcity of labeled data makes it challenging to train accurate and reliable segmentation models.

To address this issue, semi-supervised learning methods have been widely adopted in medical image segmentation, as they can leverage both labeled and unlabeled data during training. Bayesian deep learning (BDL) has emerged as a promising approach in semi-supervised medical image segmentation, as it provides a principled way to quantify uncertainty and make probabilistic predictions. However, current BDL methods for medical image segmentation are often discriminative in nature, relying heavily on labeled data for training, which can lead to overfitting due to the limited availability of labeled data in the medical domain. Moreover, these discriminative models may not generate trustworthy pseudo labels for semi-supervised training, and lack clear Bayesian formulations.

Some relevant previous works included using teacher-student models and MC dropout ([2], [3], [5], [6], [7]) where the general idea is for the student model to learn from the teacher model and improve it's segmentation performance. A type of this teacher-student model is the uncertainty-aware teacher-student model [6] which leverages MC dropout to remove untrustworthy predictions done by the teacher model and the clean predictions are used by the student model to learn from the unlabeled data.

In this paper, the authors propose a novel generative Bayesian deep learning (GBDL) architecture for semi-supervised medical image segmentation. The main goal of GBDL is to model the joint distribution of labeled images, unlabeled images, and their corresponding labels, denoted as $P(X, Y)$, where X represents the images and Y represents the labels. Unlike previous methods, the proposed GBDL approach explicitly models the joint distribution of images and labels, instead of relying solely on discriminative modeling. This allows GBDL to capture the underlying probabilistic relationships between images and labels, leading to more accurate and reliable pseudo labels and there-

fore better segmentation results.

The proposed GBDL approach incorporates Bayesian inference into the generative modeling framework, allowing the authors to obtain probabilistic estimates of the model parameters and uncertainty estimates for the predictions. This solid theoretical foundation provides the authors with more reliable and interpretable segmentation results, with well-calibrated uncertainty measures. Furthermore, the authors evaluate their proposed GBDL architecture on three medical image datasets, and experimental results demonstrate that it outperforms current methods by a significant margin in terms of segmentation accuracy. In the next section we will discuss about the architecture and how the authors formulate GBDL. It is then followed by the results which the authors report, the reproduced results by us and also a few ablation studies to test the efficiency of GBDL.

2. Methods

The proposed method in this paper utilizes a two-step learning procedure in conjunction with an inference step for GBDL. The ultimate goal is to effectively utilize both labeled and unlabeled images in the first step of the learning process, generating plausible images and corresponding ground truth for the second part of the learning process. In the paper, the authors use the generated labels only for the unlabeled images and merge them with the labeled data for training. In the following subsections, a detailed description of the two-step learning procedure and inference step employed in the GBDL architecture are provided.

2.1. Learning Procedure

The primary goal of the learning procedure is to effectively estimate the posterior distribution denoted as $P(W|X,Y)$, where W represents the weights of the learning procedure, and X and Y denote the input images and their corresponding labels, respectively. To achieve this, the learning procedure employs two architectures: the Latent Representation Learning (LRL) and the 3D-UNet with Monte Carlo (MC) dropout (3DUMC), which are utilized for learning the representations and subsequently performing the image segmentation task. The schematic diagram of the learning procedure is illustrated in Fig. 1. The whole learning procedure can be denoted mathematically by the following equation (equation from [4]):

$$P(W|X,Y_L) = \int\int\int P(W|X,Y)P(X,Y|Z)P(Z|X,Y_L)dZdXdY \quad (1)$$

Upon closer examination of the equation presented in Eq. (1), we can discern that $P(Z|X,Y_L)$ represents the objective of the Latent Representation Learning (LRL) step. Subsequently, we obtain the joint distribution denoted as $P(X,Y|Z)$ based on the latent representations Z. From this joint distribution, we can then sample out values for both X and Y using $P(X,Y)$. Finally, the prior distribution $P(W|X,Y)$, where W signifies the weights of the segmentation network, can be learned using the obtained samples of X and Y. This multi-step process allows us to effectively estimate the prior distribution of the segmentation network weights based on the latent representations and the joint distribution of the input images and their corresponding labels. We know that the Eq. (1) is intractable. For this the MC approximation of the same is taken. Generally, the MC approximation is taken by drawing random samples from the proposal distribution and the intractable equation is replaced by an average of evaluations at these sampled points. The MC approximation is given by (equation from [4]):

$$\mathrm{P(W} - \mathrm{X, Y}_L) = \frac{1}{MN}\sum_{i=0}^{N-1}\sum_{j=0}^{M-1} P(W|X_{(i,j)}, Y_{(i,j)}) \quad (2)$$

In the above equation, M represents the number of latent representations drawn from $P(Z|X,Y_L)$ and N represents pair of input volumes and labels drawn from $P(X,Y|Z)$.
In the following sections, we will discuss each step of the learning procedure in detail to provide a comprehensive understanding of the proposed approach. The overall is shown in Sec. 2.1.2

2.1.1 LRL

The main objective of the LRL is to learn the distribution of the latent representations Z from X and Y_L which are all the images (labeled + unlabeled) and the labels that is present in the dataset.
The LRL architecture comprises two sets of encoders and decoders, which we will refer to as the "upper autoencoder" and "lower autoencoder". The upper autoencoder is designed as a conditional variational autoencoder (cVAE), which aims to reconstruct the input images/volumes. It assumes a latent distribution of Z denoted as $Q(Z)$, which is modulated by the input volumes X, resulting in a conditional distribution $Q(Z|X)$. The optimal learning of representations is achieved by minimizing the distance between $Q(Z|X)$ and $P(X)$ which is the distribution of the input images. In other words the task is to minimize the KL divergence in between the two distributions. Since the upper autoencoder is focusing on reconstructing the input image, the mean squared error loss (L_{MSE}) is calculated for optimised reconstruction.
Even though the upper autoencoder follows the structure of cVAE, the authors claim that there are some crucial differences in between the cVAE and LRL. Firstly, the cVAE only generates images based on image data distribution and ignores labels Y, on the other hand, LRL considers joint distribution of image X and label Y both. Secondly, cVAE

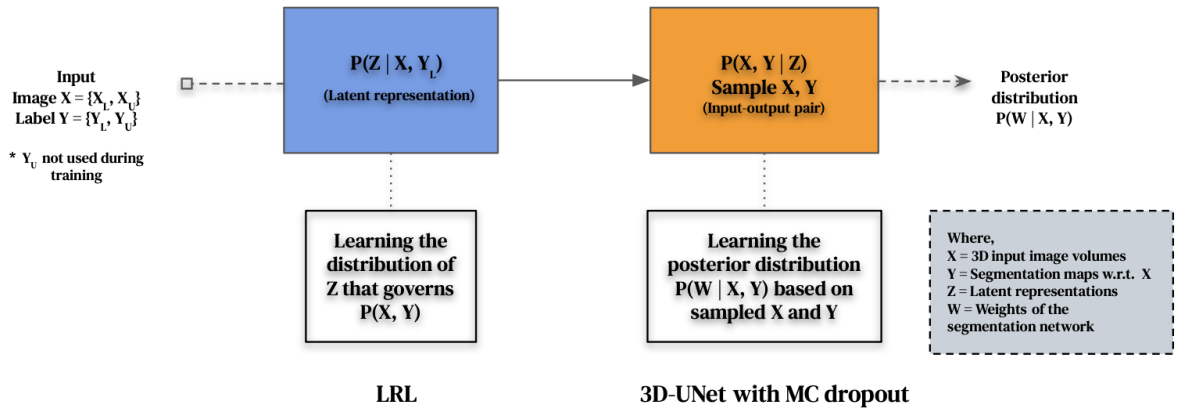

Figure 1. Schematic diagram of the learning procedure in the GBDL. The blue box depicts the first stage of the learning procedure and the orange box depicts the second stage of the learning procedure.

assumes independence and encodes mean and variance vectors for latent representations whereas LRL does not assume independence and encodes mean and variance matrices for each and every slice of the 3D image which takes us to the last difference, cVAE is generally designed for 2D inputs whereas LRL is designed for 3D inputs.

Moving on to the lower autoencoder, it is a simple autoencoder that focuses on generating segmentation maps based on the merged latent representations which is given by $P(Y_L|X_L, Z)$. The latent representation for each slice is obtained by multiplying the learned representations from the upper autoencoder with the representations obtained from the encoder of the lower autoencoder. These merged latent representations are then passed into the decoder of the lower autoencoder, along with the latent representations learned from its own encoder, to generate segmentation maps. Since the lower autoencoder focuses on reconstructing the segmentation maps, it's goal would be to minimize the cross-entropy (L_{CE}) and dice loss (L_{Dice}).

From the above explanation of the LRL architecture we can say that the learning objective of LRL is to maximize the following evidence lower bound (ELBO) (equation from [4]):

$$\log \mathrm{P(X,\,Y)} \geq \mathbb{E}_Q[logP(X|Z) + logP(Y|X,Z)] - \mathbb{E}_Q[log(\tfrac{Q(Z|X)}{P(Z)})] \quad (3)$$

where $\mathbb{E}_Q$ is expectation over $Q(Z|X)$. In the ELBO, we see that first term $\mathbb{E}_Q[logP(X|Z) + logP(Y|X,Z)]$ are basically the objectives of the upper autoencoder and lower autoencoder. As we have defined the losses above, the ELBO in terms of losses can also be written as (equation from [4]):

$$\mathrm{L}_{ELBO} = \lambda_1 L_{CE} + \lambda_2 L_{Dice} + \lambda_3 L_{MSE} + \lambda_4 L_{KL[Q(Z|X)||P(Z)]} \quad (4)$$

The above is the overall loss function of the LRL. The coefficients λ_1, λ_2, λ_3 and λ_4 are set by the authors to 1.0, 2.0, 1.0 and 0.005 by using the variable-controlling approach.

In summary, the LRL architecture in this work comprises an upper autoencoder that reconstructs input images and learns a conditional distribution of Z from X, and a lower autoencoder that generates segmentation maps based on the merged latent representations. The architecture is designed to learn meaningful representations of the data and optimize the distance between the learned and input distributions, making it a crucial component of the proposed approach.

2.1.2 3D-UNet with MC dropout (3DUMC)

The 3D-UNet with MC dropout [1] (3DUMC) serves as the second stage of the overall learning procedure in this work. After the completion of the first step, where the Latent Representation Learning (LRL) is trained, the LRL is fixed and the focus shifts to training the 3DUMC. The main goal of this second step is to learn the posterior distribution $P(W|X,Y)$, where W represents the weights of the segmentation network. To achieve this, the pseudo labels generated by the LRL, along with the labeled images, are combined as input to the 3DUMC. The 3DUMC is designed as a typical autoencoder, with both encoder and decoder architectures based on 3D Convolutional Neural Networks (CNNs). In order to incorporate MC dropout into the 3DUMC, dropout layers are added after each layer in the 3D decoder. This helps in regularizing the model and accounting for model uncertainty during training. Similar to the lower autoencoder in the LRL, the objective of the 3DUMC is to minimize the cross-entropy and dice loss, which are common loss functions used in segmentation tasks. As a result, the final loss function L_{Seg} for the 3DUMC is a com-

bination of the cross-entropy and dice loss, along with the regularization from the MC dropout layers. The final loss functions looks like the below (equation from [4]),

$$L_{Seg} = \beta_1 L_{CE} + \beta_2 L_{Dice} \tag{5}$$

The values of β_1 and β_2 are set 1.0 and 2.0 by using the variable-controlling approach. The authors use the MC dropout part of the 3DUMC only in the inference phase. We will discuss the inference phase in the next section.

2.2. Inference Procedure

The inference procedure is performed after the completion of the learning procedure. It involves using the learned weights W obtained from the last step of the learning procedure on the test images denoted as X_{test}, in order to obtain the predicted labels Y_{pred}. The Fig. 2 depicts the overall schematic diagram of the inference procedure.
The inference procedure can therefore be formulated as (equation from [4]):

$$\mathrm{P}(\mathrm{Y}_{pred}|X_{test}, X, Y_L) = (Y_{pred}|X_{test}, W)P(W|X, Y_L)dW \tag{6}$$

Similar to the learning procedure, we take the MC approximation of Eq. (6), we get the following (equation from [4]):

$$\mathrm{P}(\mathrm{Y}_{pred}|X_{test}, X, Y_L) = \frac{1}{T}\sum_{i=0}^{T-1} P(Y_{pred}|X_{test}, W_i) \tag{7}$$

where T is the models drawn from the posterior distribution $P(W|X, Y)$. To account for the MC approximation, T models are sampled from the distribution output of 3DUMC denoted as $P(W|X, Y)$, as depicted in the figure. Each time a new model W_i is sampled, a different prediction result can be obtained. The final prediction is obtained by averaging all the drawn predictions, resulting in an ensemble-based prediction that incorporates the variability introduced by the MC approximation. This allows for a more robust and reliable estimation of the predicted labels for the test images.

3. Experiments

3.1. Dataset

The experiments conducted in this paper aim to evaluate the performance of the proposed Generative Bayesian Deep Learning (GBDL) approach compared to the state-of-the-art methods on three publicly available medical image datasets, namely the Kidney Tumour Segmentation dataset, Liver Segmentation dataset, and Atrial Segmentation dataset. The Kidney Tumour Segmentation dataset and the Liver Segmentation datasets both contain CT scans with a total of 210 (160 for training and 50 for testing) and 131 (100 for training and 31 for testing) labeled scans respectively. The Atrial segmentation dataset contains the MRI scans with a total of 100 labeled images out of which 80 are used for training and 20 are used for testing. For reproducing the results of the paper, we used the Atrial segmentation dataset, the experiments and results discussion for this report would therefore be based on this dataset. The images in the Atrial segmentation dataset are 3D MRI scans. The authors divide the whole image and label into 32 slices/voxels for training and testing.

3.2. Evaluation Metrics

To assess the segmentation accuracy and reliability of the proposed approach, several evaluation metrics are utilized. These metrics include the Dice score, Jaccard score, 95% Hausdorff Distance, and Average Surface Distance. These metrics are widely used in the field of medical image segmentation as they provide comprehensive quantitative measures for evaluating the quality of segmentation results. The Dice score and Jaccard score are measures of overlap between the predicted segmentation and the ground truth, with higher scores indicating better segmentation accuracy. The 95% Hausdorff Distance is a measure of the maximum distance between the predicted and ground truth segmentations, capturing the extent of spatial dissimilarity. The Average Surface Distance measures the average distance between the predicted and ground truth segmentation along the surface of the segmented region, providing information on the spatial accuracy of the segmentation.

3.3. Comparison with state-of-the-art

In the comparison with state-of-the-art methods, the proposed Generative Bayesian Deep Learning (GBDL) approach is evaluated on the Atrial Segmentation dataset using different ratios of labeled and unlabeled data, and compared with other semi-supervised medical image segmentation methods.
Two different ratios of labeled and unlabeled data are used in the experiments: 0.2 and 0.1. In the first ratio (0.2), out of 80 samples, 16 scans are labeled and 64 scans are unlabeled, while in the second ratio (0.1), out of 80 samples, 8 scans are labeled and 72 scans are unlabeled. These ratios are chosen to investigate the performance of the proposed approach with varying levels of labeled data availability.
The results in the Fig. 3 obtained from the experiments are presented in a results table, which demonstrates that the proposed GBDL approach outperforms all of the state-of-the-art methods in both the 0.2 and 0.1 labeled-to-unlabeled data ratios. The results obtained demonstrate that GBDL outperforms all previously proposed Bayesian deep-learning-based methods. This confirms that GBDL surpasses the teacher-student architecture utilized in those methods, and underscores the effectiveness of the generative learning paradigm in generating pseudo-labels for im-

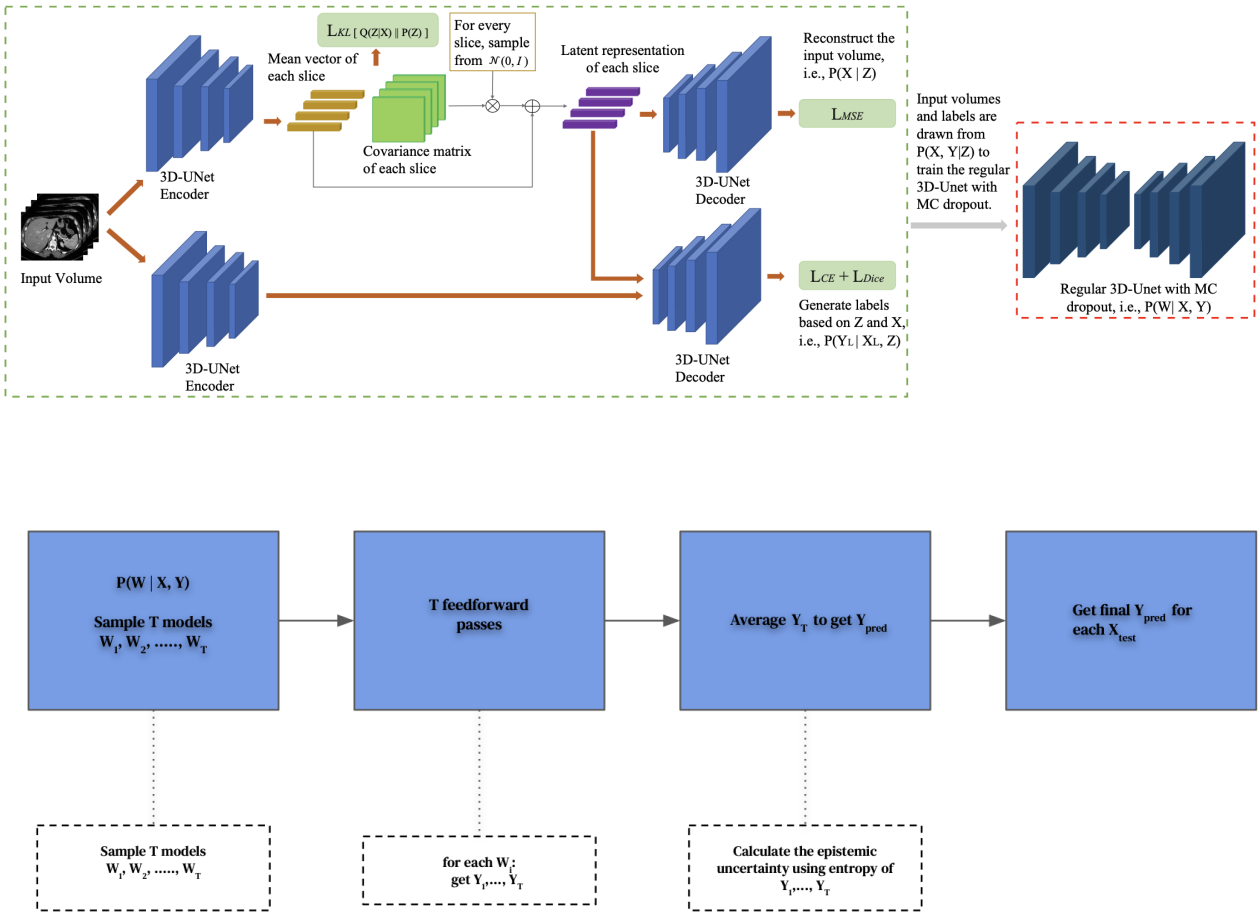

Figure 2. Schematic diagram of the inference procedure in the GBDL.

proved performance.

3.4. Reproducing the results

In order to ensure the reproducibility of the experimental results, we meticulously followed the learning procedure as discussed in the previous sections. To achieve this, we utilized the GBDL framework provided by the authors as a foundation, and made necessary modifications to the specific parts that needed to be re-implemented. The re-implementation of the Latent Representation Learning (LRL) component was carried out based on the detailed architecture information provided in the supplementary material. We made sure to adhere to the original design and settings as much as possible in order to replicate the learning procedure employed by the authors. Furthermore, our reproducibility study extended beyond just replicating the learning procedure. We also designed and conducted ablation studies to further investigate the performance of GBDL under different conditions. This comprehensive approach allowed us to thoroughly evaluate the reproducibility of the results and validate the effectiveness of GBDL in our experiments.

Metric	GBDL	GBDL (our baseline)	Reproduced
Dice score	0.894	0.834	0.804
Jaccard index	0.822	0.724	0.703
Hd95	4.03	6.08	6.24
ASD	1.48	1.91	2.19

Table 1. Reproduced results

In the original implementation, the authors reported a mean dice score of 0.894 after training the GBDL model for 240 epochs. However, due to the limited computational resources available to us, which included 4 GPUs with a memory capacity of 64GB each, it took us approximately 7 hours to complete 240 epochs of training. Considering the time constraints of our experiment, we made the

	Scans used		Metrics			
	Labeled	Unlabeled	Dice ↑	Jaccard ↑	95HD ↓	ASD ↓
UA-MT [21]	16	64	0.889	0.802	7.32	2.26
SASSNet [9]	16	64	0.895	0.812	8.24	2.20
Double-UA [18]	16	64	0.897	0.814	7.04	2.03
Tripled-UA [17]	16	64	0.893	0.810	7.42	2.21
CoraNet [15]	16	64	0.887	0.811	7.55	2.45
Reciprocal Learning [22]	16	64	**0.901**	0.820	6.70	2.13
DTC [10]	16	64	0.894	0.810	7.32	2.10
3D Graph-S^2 Net [6]	16	64	0.898	0.817	6.68	2.12
LG-ER-MT [5]	16	64	0.896	0.813	7.16	2.06
Double-UA* [18]	16	64	0.894	0.809	6.16	2.28
UA-MT* [21]	16	64	0.891	0.793	6.44	2.39
Tripled-UA* [17]	16	64	0.889	0.809	6.88	2.48
CoraNet* [15]	16	64	0.883	0.805	6.73	2.67
GBDL	16	64	0.894	**0.822**	**4.03**	**1.48**
UA-MT [21]	8	72	0.843	0.735	13.83	3.36
SASSNet [9]	8	72	0.873	0.777	9.62	2.55
Double-UA [18]	8	72	0.859	0.758	12.67	3.31
Tripled-UA [17]	8	72	0.868	0.768	10.42	2.98
CoraNet [15]	8	72	0.866	0.781	12.11	2.40
Reciprocal Learning [22]	8	72	0.862	0.760	11.23	2.66
DTC [10]	8	72	0.875	0.782	8.23	2.36
LG-ER-MT [5]	8	72	0.855	0.751	13.29	3.77
3D Graph-S^2 Net [6]	8	72	0.879	0.789	8.99	2.32
Double-UA* [18]	8	72	0.864	0.767	10.99	3.02
UA-MT* [21]	8	72	0.847	0.744	12.32	3.20
Tripled-UA* [17]	8	72	0.868	0.760	9.73	3.31
CoraNet* [15]	8	72	0.861	0.770	11.32	2.46
GBDL	8	72	**0.884**	**0.792**	**5.89**	**1.60**

Figure 3. Comparison of GBDL segmentation results with other semi-supervised state-of-the-art methods. Table credit [4]

decision to set a baseline by training the model for 100 epochs. This yielded us a dice score of 0.834. These results are shown in Tab. 1. The label to unlabeled ratio is 20:80. The table presents three sets of results: the GBDL column displays the outcomes reported in the original paper, the GBDL (baseline) column shows the results obtained when we trained the framework for 100 epochs, and the Reproduced column showcases the outcomes from our re-implementation of the learning procedure. Upon careful comparison, we observe that the reproduced results are nearly similar with the ones reported by the authors. However, it is worth noting that there may be some differences between the baseline and reproduced results, as the authors employed a complex VGG backbone for their encoders and decoders in the GBDL, while we used vanilla Conv3D layers with PyTorch, as per the details provided in the supplementary material.

3.5. MC dropout vs Model Ensembling

In the paper, the authors have employed a 3D-UNet with MC dropout as the second stage of their learning procedure. While the paper mentions the use of MC dropout for uncertainty estimation, a comprehensive explanation for choosing this specific method is lacking. Therefore, it could be beneficial to compare the results obtained with MC dropout to those achieved using other uncertainty estimation techniques, such as model ensembling. This comparison would provide a more comprehensive evaluation of the performance of the Bayesian deep learning model in terms of uncertainty estimation. We trained three different UNet models with different random seeds to create an ensemble. By averaging the predictions of these ensemble models, we obtained the metrics for evaluation. However, it is important to note that we did not report the uncertainty estimate metrics in this analysis, but rather focused on the metric reported in the original paper (as referenced in Tab. 2). Interestingly, we observed that the performance of the ensemble model was better compared to using MC dropout in terms of the reported metrics. However, it is worth mentioning that this improvement came at the cost of additional training time, as training three separate models for ensembling requires more computational resources and effort compared to using MC dropout.

3.6. Ablation Studies

The authors conducted several ablation experiments to validate their design choices in the GBDL framework. One such choice was to use 3D slices/voxels as inputs to the LRL, instead of employing complete scans/volumes as inputs. This decision resulted in a notable improvement of 0.04 points in the performance. Additionally, they performed an ablation study to demonstrate that using two separate encoders in the LRL, rather than a shared encoder for the upper and lower autoencoders, is a superior approach. This experiment revealed an improvement of 0.03 in the results when separate encoders were used, providing further evidence in support of their claim. These ablation studies provide empirical evidence for the effectiveness of the authors' design choices in the GBDL framework.

4. Discussion

In addition to the ablation studies conducted by the authors, we also performed our own set of ablations to thoroughly assess the choices and robustness of the GBDL framework. These additional ablation experiments were designed to further investigate the effectiveness of certain design choices and evaluate the performance of GBDL under different conditions. The results of these ablation studies will be discussed as a part of the following sections.

4.1. Out-of-Distribution data with GBDL

Even though the authors design strong ablations to prove their choices, the performance of GBDL in real-world scenarios, particularly when faced with out-of-distribution (OOD) data with noise or ghosting artifacts, remains a topic of investigation. Here we design a study to assess the performance of GBDL on OOD data settings. OOD data can represent data that is significantly different from the training data, and evaluating how Bayesian deep learning models perform on such data is crucial to assess their generalization capability, robustness, and reliability. We will evaluate the model's segmentation results and gain insights into the model's understanding of the data prior and potential biases.

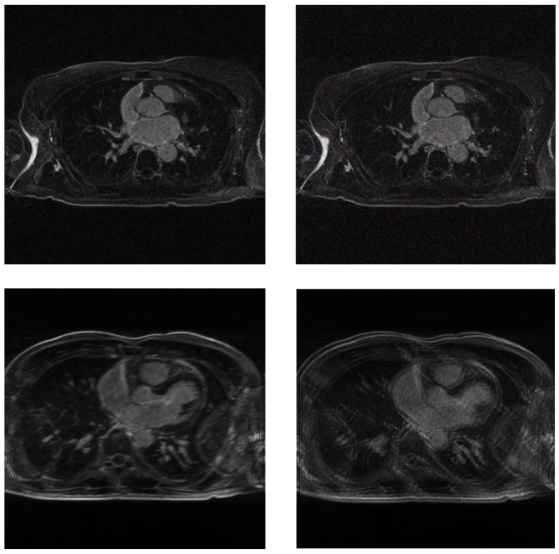

Figure 4. Types of out-of-distribution data. The upper row shows data with added gaussian noise and the lower row shows data with added ghosting atrifact

Metric	3DUMC	3DUE
Dice score	0.834	0.848
Jaccard index	0.724	0.747
Hd95	6.08	5.85
ASD	1.91	1.86

Table 2. 3D-UNet with MC dropout vs 3D-UNet ensemble

4.1.1 Noise artifact

Adding noise to MRI scans in the form of random perturbations or artifacts can simulate real-world scenarios where imaging data may be corrupted by various sources of noise, such as acquisition artifacts, hardware limitations, or patient motion. We introduced controlled levels of gaussian noise to the MRI scans to mimic the challenges that GBDL may encounter in real-world settings. This allows us to assess the model's performance in more realistic and challenging conditions, where the data may deviate from the ideal conditions assumed during training. Additionally, the introduction of noise in the MRI scans can help in evaluating the model's ability to capture and model the underlying data prior. The performance of the model on noisy MRI scans can reveal how well it can adapt to different levels of noise and whether it can accurately segment structures of interest in the presence of such noise.

We added two levels of gaussian noise to the test images with a standard deviation σ of 20 and 30. Adding more noise, in our opinion, would make the images unrealistic in nature therefore not simulate the real-world. The upper row of the Fig. 4 shows the simulated noisy images.

Metric	3DUMC	3DUMC N($\sigma = 20$)	3DUMC N($\sigma = 30$)
Dice score	0.834	0.830	0.819
Jaccard index	0.724	0.718	0.701
Hd95	6.08	6.09	6.56
ASD	1.91	1.93	2.07

Table 3. Results of GBDL on noisy data

Metric	3DUMC	3DUMC G($\theta = 20$)	3DUMC N($\theta = 30$)
Dice score	0.834	0.825	0.820
Jaccard index	0.724	0.709	0.704
Hd95	6.08	6.40	6.34
ASD	1.91	2.03	2.04

Table 4. Results of GBDL on ghosting artifact data

Tab. 3 shows inference performance of GBDL on the noisy test data demonstrate the sensitivity of the framework to different levels of noise. Upon careful examination, it can be observed that when only a minimal or negligible amount of noise ($\sigma = 20$) is added to the test set, the impact on the segmentation metrics is relatively insignificant. However, as the amount of noise increases, particularly when $\sigma = 30$, there is a visible decrease in the dice score. This suggests that GBDL is susceptible to the presence of noise artifacts, which can potentially affect its segmentation performance.

4.1.2 Ghosting artifact

The other type of effect we added to the test images for simulating OOD setting was ghosting artifact. Ghosting effects are commonly caused by the presence of motion during image acquisition, which leads to duplicate or shifted images appearing in the reconstructed MRI volume. They can introduce misalignments, duplications, or shifts in the image, which can impact the performance of the segmentation model. By evaluating the model's performance on MRI scans with ghosting effects, the study can provide insights into how the model handles such artifacts and whether it can still produce accurate segmentations despite the distortions. In this specific study, we use a ghosting shift factor θ of 20 and 30. Similar to the discussion above regarding the addition of noise, we didn't add further ghosting effect to the test data since it might render it unrealistic in nature which beats the purpose of the study.

The performance of GBDL on ghosting effect data is presented in Tab. 4. Upon careful examination, it can be observed that the addition of ghosting artifacts, as well as the

increase in θ, does not have a significant impact on the segmentation performance of GBDL. Although there is a slight decrease in the dice score by approximately 0.01, the increase in θ does not result in a substantial decrease in the overall test performance of GBDL. These findings suggest that GBDL may be relatively robust to the presence of ghosting artifacts in the medical image data, and its segmentation performance may not be severely impacted by such artifacts. However, further investigation and validation on different datasets and imaging conditions may be warranted to confirm these findings and assess the generalizability of GBDL in the presence of ghosting effects.

5. Conclusion and Future Directions

In the paper, the authors introduce a novel architecture called GBDL for semi-supervised medical image segmentation. In order to validate the claimed results, we have re-implemented the learning procedure and conducted additional experiments to assess the robustness of the proposed method. Despite the considerable training time required for GBDL, it has demonstrated superior performance compared to state-of-the-art Bayesian deep learning methods for semi-supervised image segmentation.

One of the assumptions made by the authors in GBDL is that the latent representation of each slice from the input volume, denoted as $P(Z)$, follows a multivariate Gaussian distribution. This assumption simplifies the modeling process and allows for efficient inference, but it may have limitations in capturing the true distribution, variability, and non-linearity in the data. It is important to acknowledge that the choice of a Gaussian distribution for modeling the latent representation may not always accurately represent the underlying characteristics of the data being analyzed. As a potential extension of this work, exploring alternative distributions beyond the multivariate Gaussian distribution could be considered.

Another natural extension of this work is to closely look into the uncertainty estimation and propagation in the GBDL framework, as this can provide valuable insights into the reliability and confidence of the segmentation results. This could involve investigating different techniques for uncertainty estimation, such as Model ensembling, Monte Carlo dropout, Bootstrap ensembling etc to better quantify the uncertainty associated with the predicted segmentations. This analysis could further enhance the interpretability and trustworthiness of the GBDL method in real-world medical image segmentation tasks.

As far as the model architecture is concerned, as discussed above, for the learning procedure, the LRL is trained for the first half of the epoch which is then frozen and then the 3DUMC is trained. Since there is no flow of information from the LRL to the 3DUMC, it might lead to unstable training in a few cases. In the case of erraneous pseudo-labels, the 3DUMC would get trained on bad ground truths finally leading to missegmentation and inaccurate results. In our opinion, this problem can be alleviated by introducing a feedback mechanism or iterative refinement process between the LRL and 3DUMC during training. For example, after training the LRL for the first half of the epoch, instead of freezing it completely, we could allow for partial updates of the LRL based on feedback from the 3DUMC. This would enable the LRL to adjust its predictions based on the performance of the 3DUMC, and potentially correct any errors or uncertainties in the pseudo-labels. Alternatively, we could incorporate a joint optimization approach, where the LRL and 3DUMC are trained together in an end-to-end manner, allowing for continuous exchange of information and adaptation between the two components.

Computational credits: All the experiments in this study have been conducted using computational resources from Mila - Quebec AI Institute and NeuroPoly Lab, Polytechnique Montreal.

References

[1] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, 2016. 3

[2] Suman Sedai, Bhavna Antony, Ravneet Rai, Katie Jones, Hiroshi Ishikawa, Joel Schuman, Wollstein Gadi, and Rahil Garnavi. Uncertainty guided semi-supervised segmentation of retinal layers in oct images, 2021. 1

[3] Yinghuan Shi, Jian Zhang, Tong Ling, Jiwen Lu, Yefeng Zheng, Qian Yu, Lei Qi, and Yang Gao. Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation, 2021. 1

[4] Jianfeng Wang and Thomas Lukasiewicz. Rethinking bayesian deep learning methods for semi-supervised volumetric medical image segmentation, 2022. 1, 2, 3, 4, 6

[5] Kaiping Wang, Bo Zhan, Chen Zu, Xi Wu, Jiliu Zhou, Luping Zhou, and Yan Wang. Tripled-uncertainty guided mean teacher model for semi-supervised medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II*, page 450–460, Berlin, Heidelberg, 2021. Springer-Verlag. 1

[6] Yixin Wang, Yao Zhang, Jiang Tian, Cheng Zhong, Zhongchao Shi, Yang Zhang, and Zhiqiang He. Double-uncertainty weighted method for semi-supervised learning, 2020. 1

[7] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation, 2019. 1