Rethinking Bayesian Deep Learning Methods for
Semi-Supervised Volumetric Medical Image Segmentation

Jianfeng Wang
Department of Computer Science
University of Oxford
jianfeng.wang@cs.ox.ac.uk

Thomas Lukasiewicz
Department of Computer Science
University of Oxford
thomas.lukasiewicz@cs.ox.ac.uk

Abstract

Recently, several Bayesian deep learning methods have been proposed for semi-supervised medical image segmentation. Although they have achieved promising results on medical benchmarks, some problems are still existing. Firstly, their overall architectures belong to the discriminative models, and hence, in the early stage of training, they only use labeled data for training, which might make them overfit to the labeled data. Secondly, in fact, they are only partially based on Bayesian deep learning, as their overall architectures are not designed under the Bayesian framework. However, unifying the overall architecture under the Bayesian perspective can make the architecture have a rigorous theoretical basis, so that each part of the architecture can have a clear probabilistic interpretation. Therefore, to solve the problems, we propose a new generative Bayesian deep learning (GBDL) architecture. GBDL belongs to the generative models, whose target is to estimate the joint distribution of input medical volumes and their corresponding labels. Estimating the joint distribution implicitly involves the distribution of data, so both labeled and unlabeled data can be utilized in the early stage of training, which alleviates the potential overfitting problem. Besides, GBDL is completely designed under the Bayesian framework, and thus we give its full Bayesian formulation, which lays a theoretical probabilistic foundation for our architecture. Extensive experiments show that our GBDL outperforms previous state-of-the-art methods in terms of four commonly used evaluation indicators on three public medical datasets.

1. Introduction

Deep neural networks are a powerful tool for visual learning, and they have outperformed previous methods that are based on hand-crafted features in many classical computer vision tasks. Due to the power of deep neural networks, researchers also try to use them to segment pathology re-

gions from medical images. Training segmentation networks usually relies on large labeled medical image datasets, but they are scarce and difficult to obtain, since not everyone is qualified to annotate medical data, and there is only a limited number of medical experts. Therefore, semi-supervised segmentation has become an important research direction in the area of medical computer vision.

Several methods [1, 2, 5, 6, 9–11, 14, 15, 17–19, 21, 22] have been proposed in recent years.[1] Among these works, Bayesian deep-learning-based methods [14,15,17,18,21] are closely related to this work, as they are based on Monte Carlo (MC) dropout [4], which is an approximation of Bayesian neural networks (BNNs). These methods are mainly built with the teacher-student architecture, which can be regarded as the discriminative model. Specifically, they initially build the model $P(Y|X)$ only from labeled data, and then apply the model to generate pseudo labels that are refined or rectified based on the epistemic uncertainty [7] provided by MC dropout for unlabeled ones. Then, the pseudo labels are further combined with unlabeled and labeled data to further train the overall architecture.

The works [14, 15, 17, 18, 21] mainly have two issues. Firstly, their models are built only with labeled data in the early stage of training and may overfit to them, since the quantity of labeled data is limited. Consequently, the new pseudo labels generated by the models might be inaccurate, which adversely impacts the subsequent training process. Secondly, their models are only partially based on Bayesian deep learning and are not designed under the Bayesian framework. Hence, they are unable to give a Bayesian formulation with regard to their models, lacking a solid theoretical basis. As a result, some modules in their models are empirically designed, and the functions of those modules remain unclear.

---

[1]Note that we only consider previous methods that build their architectures with 3D CNNs, since our architecture is also based on 3D CNNs. In most cases, using 3D CNNs to process medical volumetric data performs better than using 2D CNNs, as the relationships among neighboring slices are preserved in 3D CNNs. Thus, comparing our method with 2D-CNN-based methods is unfair to them.

In this work, we aim to fix these two problems by proposing a new generative model, named generative Bayesian deep learning (GBDL) architecture, as well as its corresponding full Bayesian formulation. Unlike teacher-student-based architectures, GBDL estimates the joint probability distribution $P(X, Y)$ via both labeled and unlabeled data, based on which more reliable pseudo-labels can be generated, since the generative model is more faithful about the overall data distribution $P(X)$ and the intrinsic relationship between inputs and their corresponding labels (modeled by $P(X, Y)$), alleviating the overfitting problem. Moreover, GBDL is entirely constructed under the Bayesian framework, and its related full Bayesian formulation lays a solid theoretical foundation, so that every part of GBDL has its corresponding probabilistic formulation and interpretation.

The main contributions of this paper are as follows:

- We propose a new generative Bayesian deep learning (GBDL) architecture. GBDL aims to capture the joint distribution of inputs and labels, which moderates the potential overfitting problem of the teacher-student architecture in previous works [14, 15, 17, 18, 21].

- Compared with previous methods [14, 15, 17, 18, 21], GBDL is completely designed under the Bayesian framework, and its full Bayesian formulation is also given here. Thus, it has a theoretical probabilistic foundation, and the choice of each module in GBDL and the construction of the loss functions are more mathematically grounded.

- In extensive experiments, GBDL outperforms previous methods with a healthy margin in terms of four evaluation indicators on three public medical datasets, illustrating that GBDL is a superior architecture for semi-supervised volumetric medical image segmentation.

The rest of this paper is organized as follows. In Section 2, we review related methods. Section 3 presents our GBDL architecture and its Bayesian formulation, followed by the experimental settings and results in Section 4. In Section 5, we give a summary and an outlook on future research.

The source code is available at https://github.com/JianfWang/GBDL.

2. Related Work

In this section, we briefly review related Bayesian and other deep learning methods for semi-supervised volumetric medical image segmentation in the past few years.

2.1. Bayesian Deep Learning Methods

Previous Bayesian deep learning methods simply use MC dropout as a tool to detect when and where deep models might make false predictions [14, 15, 17, 18, 21]. Specifically, an uncertainty-aware self-ensembling model [21] was proposed based on the teacher-student model and MC dropout.

In particular, a teacher model and a student model were built, where the latter learnt from the former by minimizing the segmentation loss on the labeled data and the consistency loss on all the input data. MC dropout was leveraged to filter out the unreliable predictions and to preserve the reliable ones given by the teacher model, and the refined predictions can help the student model learn from unlabeled data. Based on this uncertainty-aware teacher-student model, a double-uncertainty weighted method [18] was proposed, which takes both the prediction and the feature uncertainty into consideration for refining the predictions of the teacher model during training. Sedai *et al.* [14] used MC dropout for training the teacher model as well, but they designed a novel loss function to guide the student model by adaptively weighting regions with unreliable soft labels to improve the final segmentation performance, rather than simply removing the unreliable predictions. Shi *et al.* [15] improved the way of uncertainty estimation by designing two additional models, which are called object conservation model and object-radical model, respectively. Based on these two models, certain region masks and uncertain region masks can be obtained to improve pseudo labels and to prevent the possible error propagation during training. Since the multi-task learning can boost the performance of segmentation, and it has not been considered in previous works, Wang *et al.* [17] combined it into their architecture, and imposed the uncertainty estimation on all tasks to get a tripled-uncertainty that is further used to guide the training of their student model.

2.2. Other Deep Learning Methods

Concerning other deep learning methods proposed recently, they are mainly classified as three categories: new training and learning strategies [1, 22], shape-aware or structure-aware based methods [5, 6, 9], and consistency regularization based methods [2, 10, 11, 19].

The new training and learning strategies aim to gradually improve the quality of pseudo labels during training. For example, Bai *et al.* [1] proposed an iterative training strategy, in which the network is first trained on labeled data, and then it predicts pseudo labels for unlabeled data, which are further combined with the labeled data to train the network again. During the iterative training process, a conditional random field (CRF) [8] was used to refine the pseudo labels. Zeng *et al.* [22] proposed a reciprocal learning strategy for their teacher-student architecture. The strategy contains a feedback mechanism for the teacher network via observing how pseudo labels would affect the student, which is omitted in previous teacher-student based models.

The shape-aware or structure-aware based methods encourage the network to explore complex geometric information in medical images. For instance, Hang *et al.* [5] proposed a local and global structure-aware entropy-regularized mean teacher (LG-ER-MT) architecture for semi-supervised

segmentation. It extracts local spatial structural information by calculating inter-voxel similarities within small volumes and global geometric structure information by utilizing weighted self-information. Li *et al.* [9] introduced a shape-aware semi-supervised segmentation strategy that integrates a more flexible geometric representation into the network for boosting the performance. Huang *et al.* [6] designed a 3D Graph Shape-aware Self-ensembling Network (3D Graph-S^2Net), which is composed by a multi-task learning network and a graph-based module. The former performs the semantic segmentation task and predicts the signed distance map that encodes rich features of object shape and surface, while the latter explores co-occurrence relations and diffuse information between these two tasks.

The consistency-regularization-based methods try to add different consistency constraints for unlabeled data. In particular, Bortsova *et al.* [2] proposed a model that is implemented by a Siamese architecture with two identical branches to receive differently transformed versions of the same images. Its target is to learn the consistency under different transformations of the same input. Luo *et al.* [10] designed a novel dual-task-consistency model, whose basic idea is to encourage consistent predictions of the same input under different tasks. Xie *et al.* [19] proposed a pairwise relation-based semi-supervised (PRS2) model for gland segmentation on histology tissue images. In this model, a supervised segmentation network (S-Net) and an unsupervised pairwise relation network (PR-Net) are built. The PR-Net learns both semantic consistency and image representations from each pair of images in the feature space for improving the segmentation performance of S-Net. Luo *et al.* [11] proposed a pyramid (i.e., multi-scale) architecture that encourages the predictions of an unlabeled input at multiple scales to be consistent, which serves as a regularization for unlabeled data.

3. Methodology

In this section, we start from giving the Bayesian formulation of our GBDL architecture. Thereafter, the architecture and related loss functions are introduced in detail.

3.1. Bayesian Formulation

Learning Procedure. The objective of semi-supervised volumetric medical image segmentation is to learn a segmentation network, whose weights are denoted by W, with partially labeled data, and the Bayesian treatment of this target is to learn the posterior distribution, namely, $P(W|X, Y_L)$. Now, we use X to denote the input volumes, which contains labeled input volumes (X_L) and unlabeled input volumes (X_U), i.e., $X = \{X_L, X_U\}$, and we use $Y = \{Y_L, Y_U\}$ to denote the ground-truth labels with respect to X. Each label in Y is a voxel-wise segmentation label map that has the same shape as its corresponding input volume. Note that

Y_U represents the ground-truth labels of X_U, which are not observed in the training data.

The whole learning procedure of $P(W|X, Y_L)$ can be written as:

$$
\begin{aligned}
&P(W|X, Y_L) \\
&= \iiint P(W|X, Y)P(X, Y|Z)P(Z|X, Y_L)dZdXdY,
\end{aligned}
\tag{1}
$$

where Z is the latent representation that governs the joint distribution of X and Y, denoted by $P(X, Y|Z)$. To estimate the joint distribution, one should learn Z from X and Y_L. Concerning that Eq. (1) is intractable, we take the MC approximation of it as follows:

$$
P(W|X, Y_L) = \frac{1}{MN} \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} P(W|X_{(i,j)}, Y_{(i,j)}).
\tag{2}
$$

In this approximation, M latent representations Z are drawn from $P(Z|X, Y_L)$. Then, N pairs of input volumes and labels are obtained from $P(X, Y|Z)$, which are further used to get the posterior distribution. Thus, to generate X and Y, one should obtain the joint distribution $P(X, Y)$, i.e., estimating the distribution of Z. Overall, the learning procedure contains two steps:

- Learning the distribution of Z that governs the joint distribution $P(X, Y)$.
- Learning the posterior distribution $P(W|X, Y)$ based on X and Y sampled from $P(X, Y)$.

Inference Procedure. The inference procedure can be formulated as follows:

$$
\begin{aligned}
&P(Y_{pred}|X_{test}, X, Y_L) \\
&= \int P(Y_{pred}|X_{test}, W)P(W|X, Y_L)dW,
\end{aligned}
\tag{3}
$$

where X_{test} and Y_{pred} denote the test input and the predicted result, respectively. The posterior distribution $P(W|X, Y_L)$ is learnt based on X and Y sampled from $P(X, Y)$, and thus the posterior can also be replaced by $P(W|X, Y)$. Since Eq. (3) is intractable as well, its MC approximation can be written as:

$$
P(Y_{pred}|X_{test}, X, Y_L) = \frac{1}{T} \sum_{i=0}^{T-1} P(Y_{pred}|X_{test}, W_i),
\tag{4}
$$

where T models are drawn from $P(W|X, Y)$, which is usually implemented via MC dropout [4] with T times feedforward passes. For each model (W_i) sampled from $P(W|X, Y)$, a prediction result can be obtained, and the final prediction of X_{test} can be calculated by averaging the T results. In addition, we get the epistemic uncertainty by calculating the entropy of these predictions.
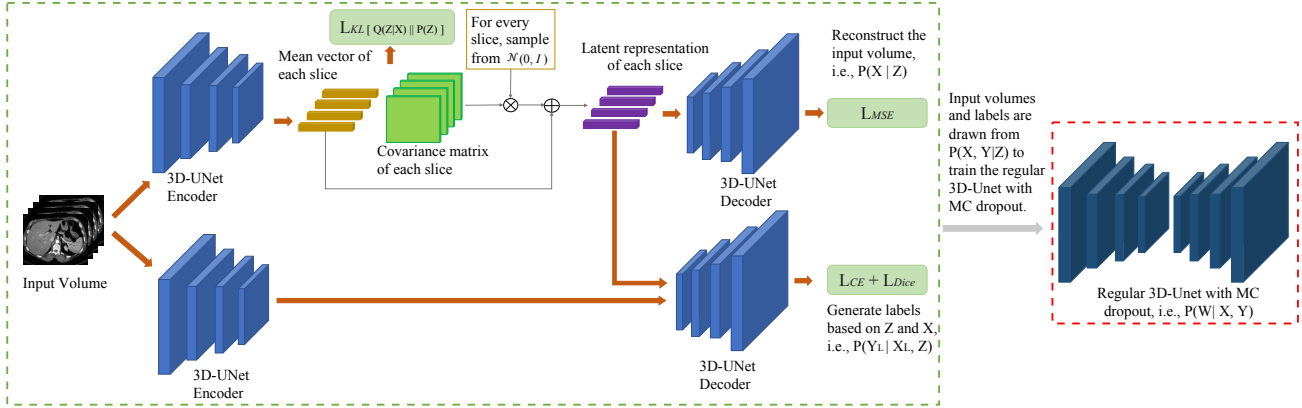
Figure 1. GBDL for semi-supervised volumetric medical image segmentation, including a latent representation learning (LRL) architecture (in the green dotted box) and a regular 3D-UNet with MC dropout (in the red dotted box). Only the regular 3D-UNet with MC dropout is used during testing. For simplicity, the shortcut connections between the paired 3D-UNet encoder and decoder are omitted.

3.2. GBDL and Loss Functions

Under the guidance of the given Bayesian formulation, we start to introduce the details of GBDL and loss functions. Concerning the two steps mentioned in the learning procedure above, GBDL has a latent representation learning (LRL) architecture and a regular 3D-UNet with MC dropout, which are shown in the green dotted box and the red dotted box in Figure 1, respectively. LRL is designed to learn the distribution of Z and to capture the joint distribution $P(X, Y)$, while the regular 3D-UNet with MC dropout is the Bayesian deep model used for parameterizing the posterior distribution $P(W|X, Y)$.

As for LRL, we first assume a variational probability distribution $Q(Z)$ to represent the distribution of the latent representation Z. In this work, we follow the conditional variational auto-encoder (cVAE) [16] to use the input to modulate Z, and thus, $Q(Z)$ can be rewritten as $Q(Z|X)$. Note that this work focuses on using 3D CNNs to process volumetric medical data, and therefore, each input volume contains several slices, $Q(Z|X)$ is actually the joint distribution of all these slices, i.e., $Q(Z|X) = q(z_0, z_1, z_2, ..., z_{n-1}|x_0, x_1, x_2, ..., x_{n-1})$, where n is the number of slices in an input volume, x_i and z_i are a slice and its latent representation, respectively. To obtain the latent representation of each slice z_i, a 3D-UNet encoder[2] is used in LRL. As the downsampling layers of the 3D-UNet encoder do not perform on the depth, and the 3D convolutional layers used in the encoder are of the kernel size 3, the padding size 1, and the stride 1 (shown in the supplementary material), each z_i is actually conditioned on n_{rf} slices, where n_{rf} is the total receptive field along the depth dimension, which

---

[2]In this paper, all 3D upsampling and downsampling layers used in the 3D-UNet encoder and the 3D-UNet decoder only perform on the spatial size, and therefore the depth of the input volumes keeps unchanged, and we can obtain the latent representation of each slice. More details about the network configuration are given in the supplementary material.

is determined by the number of 3D convolutional layers in the encoder. For ease of calculation, we assume that the distribution of the latent representation of each slice follows a multivariate Gaussian distribution and that the latent representations of slices are independent from each other. It is worth noting that such independence assumption on the latent representations of slices is reasonable, as in this way, the latent representation of each slice just rests on n_{rf} slices, which is consistent with the fact that each slice is closely relevant to its neighboring slices and distant slices may have no contribution to the slice. Therefore, based on the independence assumption, $Q(Z|X)$ can be decomposed into $\prod_{i=0}^{n-1} q(z_i|x_{(n_{rf})})$, where $x_{(n_{rf})}$ denotes the input slices that contribute to the z_i. After obtaining the latent representation Z, we can write the evidence lower bound (ELBO) as follows (with proof in the supplementary material):

$$logP(X, Y) \geq$$
$$\mathbb{E}_Q[logP(Y|X, Z) + logP(X|Z)] - \mathbb{E}_Q[log(\frac{Q(Z|X)}{P(Z)})],$$
$$(5)$$

where $\mathbb{E}_Q$ denotes the expectation over $Q(Z|X)$. Thus, the learning objective is to maximize the ELBO (Eq. (5)), which is achieved by maximizing $\mathbb{E}_Q[logP(Y|X, Z) + logP(X|Z)]$ and minimizing $\mathbb{E}_Q[log(Q(Z|X)/P(Z))]$.

Firstly, to maximize $P(X|Z)$, another 3D-UNet decoder is used to take z_i as input and the mean square error (L_{MSE}) is used as the loss function, which aims to reconstruct each slice of the input volume. Secondly, to maximize the probability $P(Y|X, Z)$, another branch can be built, which contains a 3D-UNet encoder and a 3D-UNet decoder as well. Since only parts of data have labels, $P(Y|X, Z)$ can also be rewritten as $P(Y_L|X_L, Z)$. In this branch, the 3D-UNet encoder receives X_L to extract features that are further combined with their corresponding Z to be input to the 3D-UNet decoder. Then, Y_L is leveraged to calculate the dice loss

L_{Dice} [12] and cross-entropy loss L_{CE} with X_L to maximize the probability $P(Y_L|X_L, Z)$. Thirdly, minimizing $\mathbb{E}_Q[log(Q(Z|X)/P(Z))]$ is to use $Q(Z|X)$ to approximate the prior distribution of Z, and in most cases, assuming that $Q(Z|X)$ and $P(Z)$ follow the same probability distribution can make the computation convenient. The following theorem (with proof in the supplementary material) is useful for constructing the distribution of $Q(Z|X)$.

Theorem 1. *The product of any number $n \geq 1$ of multivariate Gaussian probability density functions with precision Λ_i and mean μ_i results in an unnormalized Gaussian curve with precision $\Lambda_* = \sum_{i=0}^{n-1} \Lambda_i$ and mean $\mu_* = \Lambda_*^{-1}(\sum_{i=0}^{n-1} \Lambda_i \mu_i)$.*

Although an unnormalized Gaussian curve (denoted $f(x)$) is obtained by multiplying several multivariate Gaussian PDFs, a Gaussian distribution can still be obtained by simply normalizing $f(x)$ with $\int f(x)dx$ that does not contain x. Therefore, based on Theorem 1, we can suppose that $Q(Z|X)$ follows a multivariate Gaussian distribution. In addition, we also assume that the prior distribution $P(Z)$ follows a multivariate standard normal distribution. To minimize $\mathbb{E}_Q[log(Q(Z|X)/P(Z))]$, we propose a new loss function as follows (with proof in the supplementary material).

Corollary 1. *Minimizing $\mathbb{E}_Q[log(Q(Z|X)/P(Z))]$ is equivalent to minimizing*

$$L_{KL[Q(Z|X)||P(Z)]} =$$
$$1/2 \cdot (-log \ det[(\textstyle\sum_{i=0}^{n-1} \Lambda_i)^{-1}] + tr[(\textstyle\sum_{i=0}^{n-1} \Lambda_i)^{-1}] +$$
$$(\textstyle\sum_{i=0}^{n-1} \Lambda_i \mu_i)^T (\textstyle\sum_{i=0}^{n-1} \Lambda_i)^{-2} (\textstyle\sum_{i=0}^{n-1} \Lambda_i \mu_i) - D),$$
$$(6)$$

where n denotes the number of slices, and D denotes the number of dimensions of mean vectors.

To summarize, maximizing ELBO (Eq. (5)) is equivalent to optimizing LRL and minimizing the loss function:

$$L_{ELBO} =$$
$$\lambda_1 L_{CE} + \lambda_2 L_{Dice} + \lambda_3 L_{MSE} + \lambda_4 L_{KL[Q(Z|X)||P(Z)]},$$
$$(7)$$

where λ_1, λ_2, λ_3, and λ_4 are coefficients. Significantly, cVAE [16] has some similar properties when compared with our LRL architecture, and it is important to emphasize the differences between cVAE-based models and our LRL, which mainly contain three aspects:

- Different ELBO: cVAE is mainly used for generation tasks, in which the data distribution $P(X)$ is considered, and labels Y are not involved into the ELBO of cVAE. In contrast, our LRL concerns the joint distribution of X and Y, leading to a different ELBO (Eq. (5)).

- Independent Components Assumption: cVAE assumes that the components in the latent representation of an input

image are independent from each other. As a result, the encoder of cVAE only gives a mean vector and a variance vector for a latent representation[3]. However, our LRL does not follow this assumption for each slice, and the encoder gives a mean vector and a covariance matrix for the latent representation of each slice[4].

- Joint Variational Distribution: cVAE is mainly used for 2D inputs. In contrast, our LRL is specifically designed for 3D-CNN-based architectures for processing volumetric inputs. Therefore, since each input volume contains several slices, the variational distribution becomes a joint distribution over their latent representations, resulting in a different formulation of Kullback–Leibler (KL) divergence between the joint variational distribution and the prior distribution (Eq. (6)).

Once LRL is trained well, the joint distribution of X and Y will be captured, and the second step is to learn the posterior distribution $P(W|X, Y)$. Therefore, input volumes and corresponding labels are generated from the fixed LRL to train the regular 3D-UNet with MC dropout with two loss functions that are widely used in medical image segmentation, namely, the dice loss L_{Dice} and the cross-entropy loss L_{CE}. In this paper, we use L_{Seg} to denote their summation:

$$L_{Seg} = \beta_1 L_{CE} + \beta_2 L_{Dice}, \qquad (8)$$

where β_1 and β_2 are the coefficient of the two loss terms in L_{Seg}. In the real implementation, the new generated pseudo labels will be combined with the unlabeled data and labeled data to train the regular 3D-UNet with MC dropout. According to Eq. (4), only the the regular 3D-UNet with MC dropout is used in the test phase. In addition, as we do a full Bayesian inference during testing, it is convenient to obtain the voxel-wise epistemic uncertainty for each predicted result by calculating the entropy $(-\sum_{c=0}^{C-1} p_c log_2 p_c)$ voxel-by-voxel, where C is the number of classes.

4. Experiments

We now report on experiments of the proposed GBDL on three public medical benchmarks. To save space, implementation details are shown in the supplementary material.

4.1. Datasets

The *KiTS19* dataset[5] is a kidney tumor segmentation dataset, which has 210 labeled 3D computed tomography (CT) scans for training and validation. We followed the settings in previous works [3, 18] to use 160 CT scans for

---

[3]In practice, the encoder of cVAE gives log-variance vectors instead of variance vectors.

[4]Although the independent components assumption is widely used, it could be too strong and may degrade the performance.

[5]https://kits19.grand-challenge.org/

	KiTS19						Atrial Segmentation Challenge					
	Scans used		Metrics				Scans used		Metrics			
	Labeled	Unlabeled	Dice ↑	Jaccard ↑	95HD ↓	ASD ↓	Labeled	Unlabeled	Dice ↑	Jaccard ↑	95HD ↓	ASD ↓
3D-UNet w/ MC dropout	160	0	0.940 (±0.004)	0.888 (±0.005)	3.66 (±0.26)	0.78 (±0.11)	80	0	0.915 (±0.006)	0.832 (±0.007)	3.89 (±0.17)	1.25 (±0.14)
	16	0	0.838 (±0.044)	0.745 (±0.051)	11.36 (±1.89)	3.30 (±0.78)	16	0	0.834 (±0.044)	0.720 (±0.049)	8.97 (±1.27)	2.48 (±0.64)
	4	0	0.710 (±0.051)	0.578 (±0.065)	19.56 (±1.95)	6.11 (±0.92)	8	0	0.801 (±0.057)	0.676 (±0.051)	11.40 (±1.53)	3.27 (±0.71)
Volume-based LRL	16	144	0.892 (±0.016)	0.823 (±0.021)	7.47 (±0.39)	1.88 (±0.27)	16	64	0.871 (±0.010)	0.784 (±0.017)	5.28 (±0.37)	1.89 (±0.16)
IC-LRL	16	144	0.900 (±0.011)	0.828 (±0.012)	6.99 (±0.26)	1.75 (±0.22)	16	64	0.882 (±0.014)	0.796 (±0.026)	4.69 (±0.22)	1.66 (±0.12)
GBDL (LRL)	16	144	**0.911** (±0.010)	**0.840** (±0.013)	**6.38** (±0.37)	**1.51** (±0.22)	16	64	**0.894** (±0.011)	**0.822** (±0.011)	**4.03** (±0.41)	**1.48** (±0.14)
Volume-based LRL	4	156	0.883 (±0.011)	0.810 (±0.014)	8.32 (±0.42)	1.99 (±0.24)	8	72	0.865 (±0.017)	0.773 (±0.013)	6.81 (±0.33)	2.49 (±0.13)
IC-LRL	4	156	0.889 (±0.014)	0.814 (±0.014)	8.01 (±0.31)	2.01 (±0.12)	8	72	0.871 (±0.009)	0.779 (±0.024)	5.96 (±0.22)	1.97 (±0.17)
GBDL (LRL)	4	156	**0.898** (±0.008)	**0.821** (±0.011)	**6.85** (±0.44)	**1.78** (±0.23)	8	72	**0.884** (±0.007)	**0.792** (±0.012)	**5.89** (±0.31)	**1.60** (±0.15)

Table 1. Ablation studies of different LRL variants on the KiTS19 dataset and the Atrial Segmentation Challenge dataset. For each of our results, the "mean (±std)" is reported. The first row shows the upper bound performance, i.e., all training data with their labels are used.

	KiTS19						Atrial Segmentation Challenge					
	Scans used		Metrics				Scans used		Metrics			
	Labeled	Unlabeled	Dice ↑	Jaccard ↑	95HD ↓	ASD ↓	Labeled	Unlabeled	Dice ↑	Jaccard ↑	95HD ↓	ASD ↓
w/o L_{MSE}	16	144	0.877 (±0.014)	0.789 (±0.019)	8.89 (±0.33)	1.96 (±0.19)	16	64	0.873 (±0.012)	0.781 (±0.011)	4.76 (±0.46)	1.66 (±0.17)
w/o L_{KL}	16	144	0.868 (±0.013)	0.776 (±0.009)	8.99 (±0.29)	2.03 (±0.17)	16	64	0.876 (±0.014)	0.775 (±0.016)	5.12 (±0.55)	1.87 (±0.15)
Shared Encoder	16	144	0.854 (±0.019)	0.742 (±0.021)	10.04 (±0.67)	2.56 (±0.27)	16	64	0.862 (±0.021)	0.749 (±0.025)	6.33 (±0.73)	2.21 (±0.31)
GBDL	16	144	**0.911** (±0.010)	**0.840** (±0.013)	**6.38** (±0.37)	**1.51** (±0.22)	16	64	**0.894** (±0.008)	**0.822** (±0.011)	**4.03** (±0.41)	**1.48** (±0.14)
w/o L_{MSE}	4	156	0.864 (±0.016)	0.769 (±0.014)	9.03 (±0.39)	2.05 (±0.18)	8	72	0.867 (±0.011)	0.771 (±0.017)	4.96 (±0.34)	1.85 (±0.20)
w/o L_{KL}	4	156	0.860 (±0.017)	0.769 (±0.019)	9.32 (±0.46)	2.23 (±0.31)	8	72	0.854 (±0.015)	0.762 (±0.018)	5.33 (±0.42)	2.12 (±0.33)
Shared Encoder	4	156	0.843 (±0.021)	0.734 (±0.024)	11.08 (±0.67)	3.46 (±0.44)	8	72	0.831 (±0.019)	0.724 (±0.012)	7.19 (±0.68)	2.86 (±0.41)
GBDL	4	156	**0.898** (±0.008)	**0.821** (±0.011)	**6.85** (±0.44)	**1.78** (±0.23)	8	72	**0.884** (±0.007)	**0.792** (±0.012)	**5.89** (±0.31)	**1.60** (±0.15)

Table 2. Ablation studies of different parts in GBDL on the KiTS19 dataset and the Atrial Segmentation Challenge dataset. For each of our results, the "mean (±std)" is reported.

training and 50 CT scans for testing. The 3D scans centering at the kidney region were used, and we used the soft tissue CT window range of [-100, 250] HU for the scans.

The *Atrial Segmentation Challenge* dataset [20] includes 100 3D gadolinium-enhanced magnetic resonance imaging scans (GE-MRIs) with labels. We followed previous works [6, 9, 10, 18, 20] to use 80 samples for training and the other 20 samples for testing.

The *Liver Segmentation* dataset has been released by the MICCAI 2018 medical segmentation decathlon challenge.[6] The dataset has 131 training and 70 testing points, and the ground-truth labels of the testing data are not released. Therefore, we used those 131 CTs in our experiments, in which 100 CT scans and 31 CT scans were used for training and testing respectively. The 3D scans centering at liver regions were utilized, and we used the soft tissue CT window range of [-100, 250] HU for the scans.

4.2. Evaluation Metrics

Four evaluation metrics are used as our evaluation indicators, namely, *Dice Score*, *Jaccard Score*, *95% Hausdorff Distance (95HD)*, and *Average SurfaceDistance (ASD)*. Dice Score and Jaccard Score mainly compute the percentage of overlap between two object regions. ASD computes the average distance between the boundaries of two object regions, while 95HD measures the closest point distance between two object regions.

4.3. Ablation Studies

We conducted ablation studies on the KiTS19 and the Atrial Segmentation Challenge dataset. We comprehensively

---
[6] http://medicaldecathlon.com/index.html

searched the hyperparameters (λ_1, λ_2, λ_3, λ_4, β_1, and β_2), and the number of sampled latent representations (M) for each feedforward pass. Due to the space limit, the search process is shown in the supplementary material. In all the following experiments, we fixed λ_1, λ_2, λ_3, λ_4, β_1, β_2, and M to 1.0, 2.0, 1.0, 0.005, 1.0, 2.0, and 5 respectively.

Firstly, we evaluated the baseline model, i.e., the regular 3D-UNet with MC dropout that is trained with a limited number of labeled data, and we also evaluated the upper bound of the performance by utilizing all data to train the model. As Table 1 shows, when the number of labeled data is decreased, the performance of the regular 3D-UNet drops drastically.

Second, as for an input volume, LRL assumes that the latent representation of each slice follows a multivariate Gaussian distribution, and therefore, the latent representation of the whole volume is sampled from a joint Gaussian distribution that is obtained by multiplying the Gaussian distribution PDFs of those slices. However, LRL can also be designed to learn the latent representation of the whole volume, i.e., estimating the joint Gaussian distribution directly, and we denote this variant as "volume-based LRL". Compared to the original LRL, this variant is much closer to a 3D-CNN-based cVAE, since each input volume is treated as a whole, and the volume-based LRL represents the whole volume with a latent representation, similar to the original cVAE that represents the whole input image with a latent representation. In this case, $L_{KL[Q(Z|X)||P(Z)]}$ degenerates to $1/2 \cdot (-log \, det[\Lambda_v^{-1}] + tr[\Lambda_v^{-1}] + \mu_v^T \mu_v - D)$, where μ_v and Λ_v are the mean vector and the precision matrix of the joint Gaussian distribution, respectively. We evaluated such volume-based LRL, and the results are displayed in

	Scans used		Metrics			
	Labeled	Unlabeled	Dice ↑	Jaccard ↑	95HD ↓	ASD ↓
UA-MT [21]	16	144	0.883	0.802	9.46	2.89
SASSNet [9]	16	144	0.891	0.822	7.54	2.41
Double-UA [18]	16	144	0.895	0.828	7.42	2.16
Tripled-UA [17]	16	144	0.887	0.815	7.55	2.12
CoraNet [15]	16	144	0.898	0.820	7.23	1.89
UA-MT* [21]	16	144	0.878	0.797	8.93	2.76
Double-UA* [18]	16	144	0.899	0.823	7.55	2.21
Tripled-UA* [17]	16	144	0.886	0.819	7.43	2.26
CoraNet* [15]	16	144	0.894	0.820	7.37	1.92
GBDL	16	144	**0.911**	**0.840**	**6.38**	**1.51**
UA-MT [21]	4	156	0.871	0.787	11.74	3.56
SASSNet [9]	4	156	0.888	0.816	8.32	2.44
Double-UA [18]	4	156	0.887	0.817	8.04	2.34
Tripled-UA [17]	4	156	0.878	0.813	7.94	2.42
CoraNet [15]	4	156	0.882	0.814	8.21	2.44
Double-UA* [18]	4	156	0.890	0.819	7.93	2.33
UA-MT* [21]	4	156	0.874	0.790	11.33	3.21
Tripled-UA* [17]	4	156	0.882	0.810	7.81	2.47
CoraNet* [15]	4	156	0.886	0.812	8.43	2.39
GBDL	4	156	**0.898**	**0.821**	**6.85**	**1.78**

Table 3. Comparison with state-of-the-art semi-supervised segmentation methods on the KiTS19 dataset. "*" denotes previous methods based on the regular 3D-UNet with MC dropout.

	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-S² 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-S² 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**

Table 4. Comparison with state-of-the-art semi-supervised segmentation methods on the Atrial Segmentation Challenge. "*" denotes previous methods based on the regular 3D-UNet with MC dropout.

	Scans used		Metrics			
	Labeled	Unlabeled	Dice ↑	Jaccard ↑	95HD ↓	ASD ↓
3D-UNet w/ MC dropout	100	0	0.950	0.899	6.04	1.53
UA-MT [21]	5	95	0.920	0.867	13.21	4.54
Double-UA [18]	5	95	0.927	0.878	12.11	4.19
Tripled-UA [17]	5	95	0.921	0.869	11.77	3.62
CoraNet [15]	5	95	0.923	0.877	10.84	4.28
GBDL	5	95	**0.935**	**0.884**	**7.89**	**2.42**

Table 5. Comparison with state-of-the-art semi-supervised segmentation methods on the Liver Segmentation dataset. Previous methods in this table are also based on the regular 3D-UNet with MC dropout.

Method	KiTS19		Atrial	
	16 labeleld	4 labeleld	16 labeleld	8 labeleld
UA-MT [21]	0.868	0.841	0.851	0.836
Double-UA [18]	0.881	0.864	0.861	0.842
Triplet-UA [17]	0.874	0.852	0.864	0.847
CoraNet [15]	0.882	0.866	0.867	0.855
GBDL	**0.897**	**0.877**	**0.882**	**0.864**

Table 6. PAvPU of different Bayesian deep learning methods.

(random noise) to each slice, which is beneficial to keeping such a difference. However, the volume-based LRL only adds random noise to the latent representation of the whole volume, which goes against to keeping such a difference.

Third, considering that the independent components assumption is widely used in cVAE, we also evaluated LRL variant with this assumption, and we denote this variant as "IC-LRL". In this case, $L_{KL[Q(Z|X)||P(Z)]}$ degenerates to $1/2 \cdot \sum_D [-log \frac{1}{\sum_{i=0}^{n-1} \frac{1}{\sigma_i^2}} + \frac{1}{\sum_{i=0}^{n-1} \frac{1}{\sigma_i^2}} + (\frac{1}{\sum_{i=0}^{n-1} \frac{1}{\sigma_i^2}} \sum_{i=0}^{n-1} \frac{\mu_i}{\sigma_i^2})^2 - 1]$, where μ_i and σ_i are the mean vector and the variance vector of the Gaussian distribution for each slice in the input volume, respectively. The results shown in Table 1 indicate that the IC-LRL performs worse than our original LRL, verifying that the independent components assumption harms the performance of LRL.

Furthermore, we analyzed the two important parts of the proposed LRL, i.e., the reconstruction loss L_{MSE} and the term $L_{KL[Q(Z|X)||P(Z))]}$. We removed these two loss terms,[7] respectively, to study their impact on GBDL. As Table 2 shows, when L_{MSE} or $L_{KL[Q(Z|X)||P(Z))]}$ is removed, a performance decline can be observed, demonstrating the importance of these two terms and the correctness of our Bayesian formulation.

Besides, as the two 3D-UNet encoders in LRL are of the same architecture (see the supplementary material), and both aim at extracting features from input volumes, we evaluated a case where the two encoders share their weights. Nevertheless, sharing the weights degrades the performance, as shown in Table 2. Similarly, although some layers of the two 3D-UNet decoders in the LRL also have the same

---

[7]Removing L_{MSE} is equivalent to removing the reconstruction path, i.e., when we did an ablation study about removing the reconstruction loss L_{MSE}, we also removed the upper 3D-UNet decoder of LRL in Figure 1.
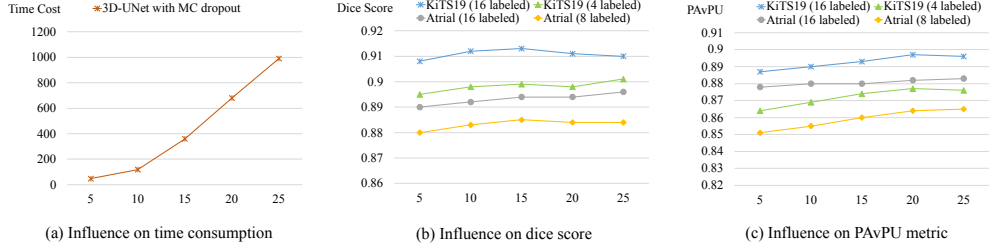
Figure 2. The influence of the number of feedforward passes on: (a) time costs (ms) for processing a $128 \times 128 \times 32$ volume input on a single GeForce GTX 1080 Ti; (b) Dice score; and (c) PAvPU metric. The horizontal axis refers to the number of feedforward passes, namely, T.

configuration, sharing their weights would clearly lead to a lower performance, as the respective objectives of them are different, leading to different gradient descent directions.

Finally, when compared with the baseline model, the proposed GBDL can improve the performance with a huge margin with respect to the four assessment metrics for both datasets. So, the proposed method is a good solution for semi-supervised volumetric medical imaging segmentation.

4.4. Comparison with State-of-the-Art Methods

The proposed GBDL is compared with previous state-of-the-art methods, and since previous works provide no standard deviation, we also only report the mean values of the four evaluation metrics in Tables 3, 4, and 5. Since most previous methods are based on VNet, for fair comparison, our regular 3D-UNet with MC dropout is designed to have similar numbers of parameters. Furthermore, we also re-implement previous Bayesian deep learning methods based on the regular 3D-UNet with MC dropout, and according to the results in Tables 3 and 4, their performance are similar to their VNet counterparts, meaning that the comparison between our method and previous methods is relatively fair.

By Tables 3, 4, and 5, the performance of GBDL is higher than previous state-of-the-art results on the three public medical datasets. When the number of labeled data decreases, the performance gap between our GBDL and previous state-of-the-art methods becomes larger, which indicates that our method is more robust when fewer labeled data were used for training. More importantly, GBDL performs better than all previous Bayesian deep-learning-based methods [15, 17, 18, 21], which verifies that GBDL is better than the teacher-student architecture used in them and that the generative learning paradigm is more suitable for pseudo-label generation and solving the semi-supervised segmentation task than the discriminative counterpart.

Furthermore, we compare the patch accuracy vs. patch uncertainty (PAvPU) metric [13] for GBDL and previous Bayesian-based methods to evaluate their uncertainty estimation. By Table 6, GBDL can output more reliable and well-calibrated uncertainty estimates than other methods, so

that clinicians can do a post-processing and refine the results in practice based on the uncertainty given by GBDL.

Finally, since full Bayesian inference is time-consuming, we show the relationship among the number of feedforward passes T, the time consumption, and the performance in Figure 2. By Figure 2(a), the time costs rise with increasing T, but by Figure 2(b) and (c), the Dice score and the PAvPU metric are not greatly impacted by T. Thus, $T = 5$ can be chosen for saving time to get the uncertainty in practice, with only a minor performance degradation.

5. Summary and Outlook

In this paper, we rethink the issues in previous Bayesian deep-learning-based methods for semi-supervised volumetric medical image segmentation, and have designed a new generative Bayesian deep learning (GBDL) architecture to solve them. The proposed method outperforms previous state-of-the-art methods on three public medical benchmarks, showing its effectiveness for handling the data with only a very limited number of annotations.

However, in GBDL, there is still a limitation with regard to the time consumption for full Bayesian inference and uncertainty estimation. By Figure 2(a), several numbers of feedforward passes will lead to a high cost in terms of time. Since Bayesian deep learning is important for some safety-critical areas, such as medical diagnosis and autonomous driving, future research can try to find better methods to improve the inference speed for GBDL. Moreover, since GBDL is a new proposed architecture, future research can also consider to combine other previous deep-learning-based methods (such as different training and learning strategies) with GBDL to further improve the performance.

References

[1] Wenjia Bai, Ozan Oktay, Matthew Sinclair, Hideaki Suzuki, Martin Rajchl, Giacomo Tarroni, Ben Glocker, Andrew King, Paul M Matthews, and Daniel Rueckert. Semi-supervised learning for network-based cardiac MR image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 253–260. Springer, 2017. 1, 2

[2] Gerda Bortsova, Florian Dubost, Laurens Hogeweg, Ioannis Katramados, and Marleen de Bruijne. Semi-supervised medical image segmentation via learning consistency under transformations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 810–818. Springer, 2019. 1, 2, 3

[3] Kang Fang and Wu-Jun Li. DMNet: Difference minimization network for semi-supervised segmentation in medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 532–541. Springer, 2020. 5

[4] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR, 2016. 1, 3

[5] Wenlong Hang, Wei Feng, Shuang Liang, Lequan Yu, Qiong Wang, Kup-Sze Choi, and Jing Qin. Local and global structure-aware entropy regularized mean teacher model for 3D left atrium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 562–571. Springer, 2020. 1, 2, 7

[6] Huimin Huang, Nan Zhou, Lanfen Lin, Hongjie Hu, Yutaro Iwamoto, Xian-Hua Han, Yen-Wei Chen, and Ruofeng Tong. 3D Graph-S^{2}Net: Shape-aware self-ensembling network for semi-supervised segmentation with bilateral graph convolution. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 416–427. Springer, 2021. 1, 2, 3, 6, 7

[7] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, 2017. 1

[8] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected CRFs with gaussian edge potentials. volume 24, pages 109–117, 2011. 2

[9] Shuailin Li, Chuyu Zhang, and Xuming He. Shape-aware semi-supervised 3D semantic segmentation for medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 552–561. Springer, 2020. 1, 2, 3, 6, 7

[10] Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8801–8809, 2021. 1, 2, 3, 6, 7

[11] Xiangde Luo, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Nianyong Chen, Guotai Wang, and Shaoting Zhang. Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 318–329. Springer, 2021. 1, 2, 3

[12] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016. 5

[13] Jishnu Mukhoti and Yarin Gal. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv:1811.12709*, 2018. 8

[14] 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. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 282–290. Springer, 2019. 1, 2

[15] 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. *IEEE Transactions on Medical Imaging*, 2021. 1, 2, 7, 8

[16] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. volume 28, pages 3483–3491, 2015. 4, 5

[17] 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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 450–460. Springer, 2021. 1, 2, 7, 8

[18] Yixin Wang, Yao Zhang, Jiang Tian, Cheng Zhong, Zhongchao Shi, Yang Zhang, and Zhiqiang He. Double-uncertainty weighted method for semi-supervised learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 542–551. Springer, 2020. 1, 2, 5, 6, 7, 8

[19] Yutong Xie, Jianpeng Zhang, Zhibin Liao, Johan Verjans, Chunhua Shen, and Yong Xia. Pairwise relation learning for semi-supervised gland segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 417–427. Springer, 2020. 1, 2, 3

[20] Zhaohan Xiong, Qing Xia, Zhiqiang Hu, Ning Huang, Cheng Bian, Yefeng Zheng, Sulaiman Vesal, Nishant Ravikumar, Andreas Maier, Xin Yang, et al. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. *Medical Image Analysis*, 67:101832, 2020. 6

[21] 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. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 605–613. Springer, 2019. 1, 2, 7, 8

[22] Xiangyun Zeng, Rian Huang, Yuming Zhong, Dong Sun, Chu Han, Di Lin, Dong Ni, and Yi Wang. Reciprocal learning for semi-supervised segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 352–361. Springer, 2021. 1, 2, 7