



This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

# A Bayesian Approach to Weakly-supervised Laparoscopic Image Segmentation

Zhou Zheng<sup>1</sup>, Yuichiro Hayashi<sup>1</sup>, Masahiro Oda<sup>2,1</sup>,  
Takayuki Kitasaka<sup>3</sup>, and Kensaku Mori<sup>1,2,5</sup>

<sup>1</sup> Graduate School of Informatics, Nagoya University, Nagoya, 464-8601, Aichi, Japan  
zzheng@mori.m.is.nagoya-u.ac.jp

kensaku@is.nagoya-u.ac.jp

<sup>2</sup> Information Technology Center, Nagoya University, Nagoya, 464-8601, Aichi, Japan

<sup>3</sup> School of Information Science, Aichi Institute of Technology,  
Toyota, 470-0392, Aichi, Japan

<sup>4</sup> Research Center of Medical Bigdata, National Institute of Informatics,  
Chiyoda-ku, 101-8430, Tokyo, Japan

**Abstract.** In this paper, we study weakly-supervised laparoscopic image segmentation with sparse annotations. We introduce a novel Bayesian deep learning approach designed to enhance both the accuracy and interpretability of the model’s segmentation, founded upon a comprehensive Bayesian framework, ensuring a robust and theoretically validated method. Our approach diverges from conventional methods that directly train using observed images and their corresponding weak annotations. Instead, we estimate the joint distribution of both images and labels given the acquired data. This facilitates the sampling of images and their high-quality pseudo-labels, enabling the training of a generalizable segmentation model. Each component of our model is expressed through probabilistic formulations, providing a coherent and interpretable structure. This probabilistic nature benefits accurate and practical learning from sparse annotations and equips our model with the ability to quantify uncertainty. Extensive evaluations with two public laparoscopic datasets demonstrated the efficacy of our method, which consistently outperformed existing methods. Furthermore, our method was adapted for scribble-supervised cardiac multi-structure segmentation, presenting competitive performance compared to previous methods. The code is available at [https://github.com/MoriLabNU/Bayesian\\_WSS](https://github.com/MoriLabNU/Bayesian_WSS).

**Keywords:** Bayesian · Weakly-supervised · Laparoscopic image · Segmentation

## 1 Introduction

Deep learning-based methods have emerged as a promising solution in laparoscopic image segmentation, which involves model training by using images and the corresponding ground truth [8,22]. However, acquiring pixel-wise annotations remains a bottleneck due to the expertise required and the time-consuming annotation process. Thus, it is highly required to explore label-efficient learning for

this task. Weakly-supervised segmentation has become an effective paradigm, which takes advantage of sparse annotations such as scribbles, diminishing the reliance on densely annotated labels. A line of promising methods has been proposed in the medical and computer vision communities. For instance, Fuentes-Hurtado et al. [3] adopted the partial cross-entropy loss (pCE) [17] to learn from the labeled pixels while ignoring the unlabeled regions for laparoscopic image segmentation. Luo et al. [13] proposed a dual-branch network and adopted a dynamically mixed pseudo-labels supervision scheme (abbreviated to DBN-DMPLS) for scribble supervision. Liu et al. [12] introduced an uncertainty-aware self-ensembling and transformation-consistency model (abbreviated to USTM) to learn from limited supervision. Yang et al. [23] presented a method comprising of a graph-model-based scheme, i.e., graph cuts [2] and a noisy learning paradigm (abbreviated to GMBM-DLM) for weakly-supervised instrument segmentation. Some studies [18,14] investigated penalization terms to regularize training.

Despite the commendable progress made by these methods, they face an array of challenges: (i) the loss of valuable image information, (ii) the error propagation due to generated low-quality supervision signals, and (iii) the limited interpretability and uncertainty quantification. For instance, the study of [3] is based on the pCE loss [17], only focusing on the labeled pixels while ignoring the unlabeled regions. This selective attention might result in the model overlooking valuable information during training. Some schemes like DBN-DMPLS [13] and USTM [12] adopt pseudo-label strategies or consistency learning paradigms to generate additional supervision signals. However, the reliability of the generated supervision signals is heavily based on model performance, potentially propagating errors and leading to degraded accuracy. Similarly, GMBM-DLM [23] employs graph cuts to generate pseudo-labels for noisy learning. However, these pseudo-labels, originating from graph cuts, often need to be improved, still carrying a risk of accumulating training errors. Methods such as the combination of the pCE loss with the DenseCRF loss [18], and the pairing of the pCE loss with the GridCRF loss [14], integrate a Conditional Random Field (CRF) [10] term within the loss to model the conditional distribution  $p(\mathbf{y}|\mathbf{x})$ , where  $\mathbf{x}$  and  $\mathbf{y}$  are images and labels, assuming that  $\mathbf{y}$  follows a Gibbs distribution. However, these methods do not fully exploit the advantages of a joint probabilistic modeling  $p(\mathbf{x}, \mathbf{y})$ , potentially missing out on the richer representation and uncertainty quantification that a fully Bayesian approach offers. Additionally, most existing methods, including those above, do not offer uncertainty estimation for predictions. Given that such models are trained under sparse supervision, assessing the uncertainty of model outputs is essential.

Driven by these perspectives above, we propose a practical and fully Bayesian learning approach for weakly-supervised laparoscopic image segmentation. It is worth mentioning that while the work [21], which inspired our study, presented a fully Bayesian learning method for semi-supervised medical segmentation, utilizing scarce annotations, our research aims to investigate weakly-supervised segmentation. We leverage sparse annotations, leading to a distinct Bayesian formulation tailored to this specific type of supervision. Unlike existing

weakly-supervised methods, our method models the joint probability distribution  $p(\mathbf{x}, \mathbf{y})$ . By harnessing this joint distribution, our method can generate superior-quality pseudo-labels by accommodating the uncertainties present in these pseudo-labels, thereby reducing the error propagation during training. Besides, our method inherently provides uncertainty estimation for its predictions. Our method provides a more principled approach to handling sparse annotations and enhances the interpretability and reliability of the segmentation results.

Our contributions are mainly threefold: (1) we pioneer rethinking weakly-supervised laparoscopic image segmentation in a Bayesian perspective and propose a novel Bayesian deep learning method for this task, which has a theoretical probabilistic foundation and enhances the accuracy and interpretability of the segmentation results; (2) we extensively validate our method on two public datasets, CholecSeg8k [8] and AutoLaparo [22] and demonstrate its potential solution for this task; and (3) we extend our method to scribble-supervised cardiac multi-structure segmentation [1,20] and show its potential for versatility and applicability across different imaging modalities.

## 2 Methodology

### 2.1 Probabilistic modeling

**Learning stage.** Generally, given a dataset  $\mathcal{D} = \{\mathbf{x}, \mathbf{y}\}$  containing images  $\mathbf{x}$  and the corresponding ground truth  $\mathbf{y}$ , we can train a model  $\mathbf{w}$  in a fully-supervised manner. This procedure also calls modeling posterior distribution  $p(\mathbf{w}|\mathbf{x}, \mathbf{y})$  under the Bayesian framework. However, the dense annotations  $\mathbf{y}$  can not be reached in weakly-supervised segmentation. Instead, the weak annotations  $\mathbf{y}^s$  are provided. The goal is to learn the model  $\mathbf{w}$  from the degraded dataset  $\mathcal{D}^s = \{\mathbf{x}, \mathbf{y}^s\}$ , i.e., modeling  $p(\mathbf{w}|\mathbf{x}, \mathbf{y}^s)$ , represented as

$$p(\mathbf{w}|\mathbf{x}, \mathbf{y}^s) = \iiint p(\mathbf{w}|\mathbf{x}, \mathbf{y}) p(\mathbf{x}, \mathbf{y}|\mathbf{z}) p(\mathbf{z}|\mathbf{x}, \mathbf{y}^s) d\mathbf{x} d\mathbf{y} d\mathbf{z}, \quad (1)$$

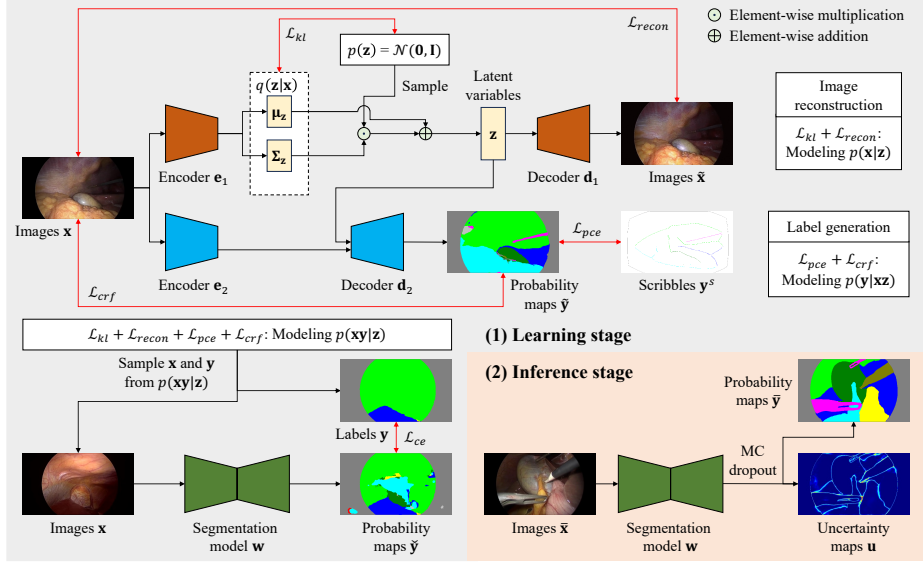
where  $\mathbf{z}$  are latent variables that determine the joint distribution  $p(\mathbf{x}, \mathbf{y})$ , and follow the posterior distribution  $p(\mathbf{z}|\mathbf{x}, \mathbf{y}^s)$ . Given the intractable nature of Eq. 1, we turn to a Monte Carlo (MC) strategy to approximate  $p(\mathbf{w}|\mathbf{x}, \mathbf{y}^s)$ :

$$p(\mathbf{w}|\mathbf{x}, \mathbf{y}^s) = \frac{1}{MN} \sum_{i=1}^M p(\mathbf{w}|\mathbf{x}_{(i)}, \mathbf{y}_{(i)}) \sum_{j=1}^N p(\mathbf{x}, \mathbf{y}|\mathbf{z}_{(j)}). \quad (2)$$

Specifically, we can first sample  $\mathbf{z}$  from  $p(\mathbf{z}|\mathbf{x}, \mathbf{y}^s)$  with  $N$  times and then draw  $M$  images and the corresponding labels from  $p(\mathbf{x}, \mathbf{y}|\mathbf{z})$  to train the model to obtain approximated  $p(\mathbf{w}|\mathbf{x}, \mathbf{y}^s)$ . Thus, the goal of the learning stage is to obtain  $p(\mathbf{z}|\mathbf{x}, \mathbf{y}^s)$  and  $p(\mathbf{x}, \mathbf{y}|\mathbf{z})$ .

**Inference stage.** After learning  $p(\mathbf{w}|\mathbf{x}, \mathbf{y}^s)$ , given test images  $\bar{\mathbf{x}}$ , we can get the corresponding probability maps  $\bar{\mathbf{y}}$  with the following formula:

$$p(\bar{\mathbf{y}}|\bar{\mathbf{x}}, \mathbf{x}, \mathbf{y}^s) = \int p(\bar{\mathbf{y}}|\bar{\mathbf{x}}, \mathbf{w}) p(\mathbf{w}|\mathbf{x}, \mathbf{y}^s) d\mathbf{w}. \quad (3)$$



**Fig. 1.** Flowchart of the proposed framework. At the learning stage, we first learn  $p(\mathbf{x}, \mathbf{y}|\mathbf{z})$  by modeling  $p(\mathbf{x}|\mathbf{z})$  for image reconstruction and  $p(\mathbf{y}|\mathbf{x}, \mathbf{z})$  for label generation. After obtaining  $p(\mathbf{x}, \mathbf{y}|\mathbf{z})$ , we sample pairs of  $\mathbf{x}$  and  $\mathbf{y}$  from  $p(\mathbf{x}, \mathbf{y}|\mathbf{z})$  to learn a segmentation model, i.e.,  $p(\mathbf{w}|\mathbf{x}, \mathbf{y})$ . At the inference stage, we obtain the prediction and corresponding epistemic uncertainty estimation with MC dropout.

We approximately calculate  $p(\bar{\mathbf{y}}|\bar{\mathbf{x}}, \mathbf{x}, \mathbf{y}^s)$  with MC simulation:

$$p(\bar{\mathbf{y}}|\bar{\mathbf{x}}, \mathbf{x}, \mathbf{y}^s) = \frac{1}{T} \sum_{i=1}^T p(\bar{\mathbf{y}}|\bar{\mathbf{x}}, \mathbf{w}_{(i)}), \quad (4)$$

where  $T$  models are sampled from  $p(\mathbf{w}|\mathbf{x}, \mathbf{y}^s)$  via MC dropout (MCDP) [4] with  $T$  inference times. By averaging all the probability maps, we can get  $\bar{\mathbf{y}}$  and the related epistemic uncertainty maps  $\mathbf{u}$  by calculating the entropy with  $-\sum_{c=1}^C p(\bar{\mathbf{y}}_c|\bar{\mathbf{x}}, \mathbf{x}, \mathbf{y}^s) \log(p(\bar{\mathbf{y}}_c|\bar{\mathbf{x}}, \mathbf{x}, \mathbf{y}^s))$ , where  $C$  is the number of classes, and  $\bar{\mathbf{y}}_c$  denotes the  $c$ -th channel of  $\bar{\mathbf{y}}$ .

## 2.2 Overview of proposed framework

The flowchart of our framework is shown in Fig. 1. The learning and inference procedures are guided by the formulation presented in Eq. 1 and 3.

We begin with the premise that  $\mathbf{z}$  are statistically independent from  $\mathbf{x}$  and  $\mathbf{y}^s$ , allowing us to simplify  $p(\mathbf{z}|\mathbf{x}, \mathbf{y}^s)$  to the prior distribution  $p(\mathbf{z})$ . This assumption sets the stage for the initial step of our method, which involves the derivation of the Evidence Lower Bound (ELBO) as shown in Eq. 5, with its proof provided in the supplementary material. Eq. 5 facilitates the decomposition of our

target log-likelihood  $\log p(\mathbf{x}, \mathbf{y})$  into manageable components, allowing us to introduce a variational distribution  $q(\mathbf{z}|\mathbf{x})$  that approximates the intractable prior distribution  $p(\mathbf{z})$ . Our objective is to maximize the ELBO, expressed as in Eq. 6.

$$\log p(\mathbf{x}, \mathbf{y}) \geq \mathbb{E}_{\mathbf{z} \sim q} [\log p(\mathbf{y}|\mathbf{x}, \mathbf{z}) + \log p(\mathbf{x}|\mathbf{z})] - \mathbb{E}_{\mathbf{z} \sim q} \left[ \log \frac{q(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})} \right]. \quad (5)$$

$$\text{ELBO} = \mathbb{E}_{\mathbf{z} \sim q} [\log p(\mathbf{y}|\mathbf{x}, \mathbf{z})] + \mathbb{E}_{\mathbf{z} \sim q} [\log p(\mathbf{x}|\mathbf{z})] - \text{KL} [q(\mathbf{z}|\mathbf{x}) || p(\mathbf{z})]. \quad (6)$$

**Latent space modeling.** We follow the conditional variational auto-encoder (CVAE) [16] to modulate the latent variables  $\mathbf{z}$  with the input images  $\mathbf{x}$ , formalizing the distribution  $q(\mathbf{z}|\mathbf{x})$ . Specifically, an encoder  $\mathbf{e}_1$  is applied to map  $\mathbf{x}$  to the latent space. Following the common settings, we assume  $q(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}}, \boldsymbol{\Sigma}_{\mathbf{z}})$  is a multivariate Gaussian distribution, and  $p(\mathbf{z})$  is a multivariate standard normal distribution  $\mathcal{N}(\mathbf{0}, \mathbf{I})$ . We calculate a Kullback–Leibler (KL) divergence loss  $\mathcal{L}_{kl}(q(\mathbf{z}|\mathbf{x}), p(\mathbf{z}))$  to push  $q(\mathbf{z}|\mathbf{x})$  closer to  $\mathcal{N}(\mathbf{0}, \mathbf{I})$ . On the other hand, a decoder  $\mathbf{d}_1$  is employed to reconstruct the images  $\tilde{\mathbf{x}}$  from the sampled latent representation<sup>5</sup>, creating a cycle that optimizes the reconstruction likelihood  $p(\mathbf{x}|\mathbf{z})$  with a loss  $\mathcal{L}_{recon}(\tilde{\mathbf{x}}, \mathbf{x})$  that calculates the mean square error (MSE).

**Conditional random field modeling.** We adopt the CRF [10], which is characterized by a Gibbs distribution, to maximize  $p(\mathbf{y}|\mathbf{x}, \mathbf{z})$ . Numerous studies like [18,14] have explored the integration of CRF into training. We use the pairing of the pCE loss [17] with the DenseCRF loss [18] to model CRF.

In practice, we introduce another encoder  $\mathbf{e}_2$  to encode the images  $\mathbf{x}$  to high-level features and concatenate them with the latent variables  $\mathbf{z}$  together and use another decoder  $\mathbf{d}_2$  to obtain the predicted probability maps  $\tilde{\mathbf{y}}$ . We make use of the sparse annotations  $\mathbf{y}^s$  to optimize  $\mathbf{e}_2$  and  $\mathbf{d}_2$  to generate sub-optimal predictions by calculating the pCE loss  $\mathcal{L}_{pce}(\tilde{\mathbf{y}}, \mathbf{y}^s)$  between  $\tilde{\mathbf{y}}$  and  $\mathbf{y}^s$ :

$$\mathcal{L}_{pce}(\tilde{\mathbf{y}}, \mathbf{y}^s) = - \sum_{a \in \Omega^s} \sum_{c=1}^C \mathbf{y}_{a,c}^s \log(\tilde{\mathbf{y}}_{a,c}), \quad (7)$$

where  $\Omega^s$  represents the set of indices corresponding to the pixels with sparse annotations. Meanwhile, we incorporate the DenseCRF loss  $\mathcal{L}_{crf}(\hat{\mathbf{y}})$ :

$$\mathcal{L}_{crf}(\hat{\mathbf{y}}) = \sum_{c=1}^C \hat{\mathbf{y}}'_c \mathbf{K}(\mathbf{1} - \hat{\mathbf{y}}_c), \quad (8)$$

where  $\hat{\mathbf{y}}_c$  is a vector associated with class  $c$ , containing all elements  $\tilde{\mathbf{y}}_{a,c}$  from  $\tilde{\mathbf{y}}_c$  for  $a \in \Omega$ , where  $\Omega$  denotes the set of indices of all pixels in  $\tilde{\mathbf{y}}_c$ , and  $\tilde{\mathbf{y}}_{a,c}$  is the component of the  $a$ -th pixel in the  $c$ -th channel of  $\tilde{\mathbf{y}}$ . Besides,  $\mathbf{K}$  is a matrix of pairwise discontinuity costs. Each element  $k_{a,b}$  in  $\mathbf{K}$  is determined by a dense Gaussian kernel [18]. By optimizing both  $\mathcal{L}_{pce}(\tilde{\mathbf{y}}, \mathbf{y}^s)$  and  $\mathcal{L}_{crf}(\hat{\mathbf{y}})$  to

<sup>5</sup> In practice,  $\mathbf{z}$  undergoes several necessary transformations between  $\mathbf{e}_1$  and  $\mathbf{d}_1$ . Details of the network configuration for this part are given in the supplementary material.

maximize  $p(\mathbf{y}|\mathbf{x}, \mathbf{z})$ , we can generate high-quality pseudo-labels and treat them as unobserved labels  $\mathbf{y}$ .

**Training procedures.** Firstly, maximizing ELBO in Eq. 6 is equivalent to optimizing the following training objective:

$$\mathcal{L}_{ELBO} = \mathcal{L}_{pce}(\tilde{\mathbf{y}}, \mathbf{y}^s) + \alpha \mathcal{L}_{kl}(q(\mathbf{z}|\mathbf{x}), p(\mathbf{z})) + \beta \mathcal{L}_{recon}(\tilde{\mathbf{x}}, \mathbf{x}) + \gamma \mathcal{L}_{crf}(\hat{\mathbf{y}}), \quad (9)$$

where  $\alpha$ ,  $\beta$ , and  $\gamma$  are weighting coefficients. It is important to note that while our work shares the same ELBO as the work [21], our method’s optimization target and loss function are distinct. The main difference is that our method assumes  $p(\mathbf{y}|\mathbf{x}, \mathbf{z})$  follows a Gibbs distribution and maximizes it with the CRF.

By optimizing  $\mathcal{L}_{ELBO}$ , we can obtain the joint distribution  $p(\mathbf{x}, \mathbf{y})$  and draw pairs of  $\mathbf{x}$  and  $\mathbf{y}$ . In the practical implementation, to make use of the scribble annotations  $\mathbf{y}^s$ , we merge  $\mathbf{y}^s$  into the generated labels  $\mathbf{y}$ . Concretely, we create binary masks  $\mathbf{\Gamma}$  based on  $\mathbf{y}^s$ , where the value of 0 indicates the labeled region, and 1 represents the unlabeled region in  $\mathbf{y}^s$ . The final labels,  $\mathbf{y}$ , are then generated by  $\mathbf{y} = (1 - \mathbf{\Gamma}) \odot \mathbf{y}^s + \mathbf{\Gamma} \odot \mathbf{y}$ , where  $\odot$  denotes element-wise multiplication. We utilize  $\mathbf{x}$  and  $\mathbf{y}$  to train a model  $\mathbf{w}$  with a cross-entropy (CE) loss  $\mathcal{L}_{ce}(\tilde{\mathbf{y}}, \mathbf{y})$  between the predicted probability maps  $\tilde{\mathbf{y}}$  and labels  $\mathbf{y}$ .

**Inference stage.** Upon completing the training of  $\mathbf{w}$ , for the test images  $\bar{\mathbf{x}}$ , we utilize MCDP [4] to obtain the probability maps  $\bar{\mathbf{y}}$  and the corresponding epistemic uncertainty maps  $\mathbf{u}$  with  $T$  inference times.

### 3 Experiments and results

#### 3.1 Experimental setup

**Datasets.** Our method was validated using two public laparoscopic datasets: CholecSeg8k [8] and AutoLaparo [22]. The CholecSeg8k dataset [8] collects 8080 images from the public dataset Cholec80 [19] and provides the corresponding ground truth of 13 classes. The resolution of each image is  $854 \times 480$  pixels. The AutoLaparo dataset [22] offers three tasks. We focused on Task 3 of segmentation. This task includes 1800 frames annotated across 10 classes. Each frame is in  $1920 \times 1080$  pixels. The dataset was split into training, validation, and test sets with 1020, 342, and 438 frames, respectively, following the official division.

**Weak label generation.** The CholecSeg8k and AutoLaparo datasets do not provide weak annotations. Inspired by previous works [3,20,23,5] that extracted skeletons from the ground truth to generate weak annotations, we obtained sparse annotations with the skeletonization method [24] in our study.

**Evaluation metrics.** We employed the Dice score (DC) [%], Jaccard (JA) [%], sensitivity (SE) [%], and specificity (SP) [%] as metrics.

**Implementation details.** We leveraged the U-Net [15] augmented with dropout layers as the backbone. Specifically,  $\mathbf{w}$  adopted the U-Net backbone. The elements  $\mathbf{e}_1$  and  $\mathbf{e}_2$  functioned as variants of U-Net’s encoder, while  $\mathbf{d}_1$  and  $\mathbf{d}_2$  were derivatives of U-Net’s decoder. The dimension of  $\mathbf{z}$  was set to 256. For hyper-parameters in Eq. 9, we empirically determined a combination of  $\alpha = 10^{-3}$ ,

**Table 1.** Quantitative comparison of various methods on the CholecSeg8k dataset and AutoLaparo test set. Experiments were conducted with 5-fold cross-validation on the CholecSeg8k dataset and repeated in 5 trials on the AutoLaparo dataset.

Method	CholecSeg8k				AutoLaparo			
	DC	JA	SE	SP	DC	JA	SE	SP
pCE [17]	51.6	46.5	53.4	98.9	25.9	20.9	24.9	94.4
DBN-DMPLS [13]	61.3	56.1	62.2	98.8	28.3	22.9	27.0	94.6
USTM [12]	64.9	58.7	66.3	99.1	26.5	21.3	25.2	94.4
GMBM-DLM [23]	78.5	71.7	76.9	99.0	28.0	23.2	24.8	94.1
pCE+DenseCRF [18]	76.6	70.4	76.8	<b>99.4</b>	28.2	22.9	25.8	94.5
pCE+DenseCRF <sup>†</sup> [18]	78.8	72.3	78.5	<b>99.4</b>	29.6	24.0	26.7	94.2
Ours w/o MCDP	80.9	74.3	80.6	<b>99.4</b>	32.1	26.2	28.8	94.6
Ours	<b>82.3</b>	<b>75.9</b>	<b>82.2</b>	<b>99.4</b>	<b>33.4</b>	<b>27.5</b>	<b>29.9</b>	<b>94.9</b>
Fully-sup	88.7	83.7	88.8	99.6	37.4	32.2	36.2	95.9

$\beta = 10^{-1}$ , and  $\gamma = 10^{-8}$ . We set  $N$  to 5 and  $T$  to 15 (see section 3.3). More details on implementation are given in the supplementary material.

**Baselines.** The upper-bound result was obtained with fully supervised segmentation (denoted as Fully-sup), while the lower-bound performance was yielded by training the model with sparse annotations via the pCE loss [17]. We further implemented four state-of-the-art (SOTA) methods for comparison: DBN-DMPLS [13], USTM [12], GMBM-DLM [23], and a combination of the pCE loss with DenseCRF losses (notated as pCE+DenseCRF) [18]. Evaluations were conducted using 5-fold cross-validation on the CholecSeg8k dataset and repeated in 5 trials on the AutoLaparo dataset.

### 3.2 Experimental results

Table 1 presents quantitative comparisons of various methods on the CholecSeg8k dataset and AutoLaparo test set. All methods performed better than the pCE loss, demonstrating their efficacy in learning from weak labels. Notably, our method consistently surpassed other SOTAs by a large margin, reaching closer to the upper-bound accuracy, particularly in DC, JA, and SE.

In addition, one might consider a straightforward approach, referred to as pCE+DenseCRF<sup>†6</sup>, which bears similarity to our approach but omits the latent variables  $\mathbf{z}$ . However, pCE+DenseCRF<sup>†</sup> fails to capture the uncertainty in the pseudo-label generation process. In contrast, by conditioning on  $\mathbf{z}$ , our method leverages sampling of  $\mathbf{z}$  to incorporate uncertainty into the model of  $p(\mathbf{x}, \mathbf{y})$ , thereby improving the pseudo-label quality in practice. To prove this, we compared our method with pCE+DenseCRF<sup>†</sup>. As shown in Table 1, while pCE+DenseCRF<sup>†</sup> generally improved the accuracy over pCE+DenseCRF, it still fell short of the performance achieved by ours with and without MCDP, underscoring the superiority of our method in generating superior pseudo-labels.

<sup>6</sup> pCE+DenseCRF<sup>†</sup> initially trains a model with pCE+DenseCRF by modeling the conditional distribution  $p(\mathbf{y}|\mathbf{x})$ , then uses this model to generate pseudo-labels and treat them as unobserved labels  $\mathbf{y}$ , and finally retrain a new model with  $\mathbf{x}$  and  $\mathbf{y}$ .

**Table 2.** Ablation studies on efficacy of loss components, influence of sample time  $N$ , and impact of inference time  $T$  with the CholecSeg8k dataset.

Loss component				Sample time	Inference time	Metrics			
$\mathcal{L}_{pce}$	$\mathcal{L}_{kd}$	$\mathcal{L}_{recon}$	$\mathcal{L}_{crf}$	$N$	$T$	DC	JA	SE	SP
✓				1	1	52.7	47.0	55.1	98.8
✓	✓			1	1	57.3	51.5	58.9	99.0
✓	✓	✓		1	1	57.9	52.0	60.0	98.9
✓	✓	✓	✓	1	1	<b>78.9</b>	<b>72.6</b>	<b>78.6</b>	<b>99.4</b>
✓	✓	✓	✓	1	1	78.9	72.6	78.6	<b>99.4</b>
✓	✓	✓	✓	3	1	<b>80.9</b>	<b>74.3</b>	<b>80.6</b>	<b>99.4</b>
✓	✓	✓	✓	5	1	79.9	73.4	80.0	<b>99.4</b>
✓	✓	✓	✓	7	1	79.0	72.6	78.8	<b>99.4</b>
✓	✓	✓	✓	3	1	80.9	74.3	80.6	<b>99.4</b>
✓	✓	✓	✓	3	5	82.1	75.7	82.0	<b>99.4</b>
✓	✓	✓	✓	3	10	82.2	75.8	82.1	<b>99.4</b>
✓	✓	✓	✓	3	15	<b>82.3</b>	<b>75.9</b>	<b>82.2</b>	<b>99.4</b>
✓	✓	✓	✓	3	20	82.2	<b>75.9</b>	<b>82.2</b>	<b>99.4</b>

**Table 3.** Quantitative comparison of segmentation performance on the ACDC dataset. Other results are adopted from [13].

Method	DC	95HD
pCE [17]	68.6	173.3
RW [6]	78.8	10.0
USTM [12]	78.6	102.2
S2L [11]	83.2	38.9
MLoss [9]	83.9	27.7
EM [7]	84.6	39.0
RLoss [17]	85.6	<b>6.9</b>
DBN-DMPLS [13]	87.2	9.9
Ours	<b>87.5</b>	9.0
Fully-sup	89.8	7.0

### 3.3 Ablation studies

Ablation studies were conducted using the CholecSeg8k dataset. These studies were divided into three stages. Initially, we evaluated each loss component’s contribution by adding them incrementally with fixed sample time  $N = 1$  and inference time  $T = 1$  (without MCDP). Next, keeping all loss components and fixing  $T = 1$  (without MCDP), we identified the optimal  $N$ . Finally, maintaining all loss components and the optimal  $N$ , we activated MCDP to assess the effects of different inference times  $T$  ( $T > 1$ ). Detailed results are presented in Table 2.

**Efficacy of loss components.** Using only  $\mathcal{L}_{pce}$  yielded the lowest performance. Sequentially adding  $\mathcal{L}_{kd}$ ,  $\mathcal{L}_{recon}$ , and  $\mathcal{L}_{crf}$  improved results, confirming each loss’s contribution. Notably, including  $\mathcal{L}_{recon}$  largely enhanced accuracy, underscoring its importance in generating high-quality pseudo-labels.

**Influence of sample time  $N$ .**  $N = 3$  obtained the best results compared to other tested values (1, 5, and 7), leading us to set  $N$  to 3 for our experiments.

**Influence of inference time  $T$ .** By sampling model weights using MCDP at 5, 10, 15, and 20 times, we found that MCDP generally improved the accuracy. As suggested by the results, we set  $T$  to 15 for all experiments.

### 3.4 Extension to other domains

We further explored the generalizability of our method by applying it to cardiac multi-structure segmentation using the ACDC dataset [1] and corresponding scribble annotations [20]. This dataset includes 200 cine-MRI volumes from 100 patients, along with the ground truth for the right ventricle, myocardium, and left ventricle. We adopted the 2D U-Net model [15] as the backbone. The implementation details are presented in the supplementary material.

Quantitative results of DC and 95% Hausdorff distance (95HD) [mm] are summarized in Table 3. We referenced results of existing methods reported in [13] for comparison purposes, considering that the same U-Net backbone and



5-fold cross-validation splitting were used. Our method showed competitive results compared to previous models, underlining its potential generalizability to different medical imaging domains.

## 4 Discussion and conclusions

We proposed a novel method grounded in a fully Bayesian learning paradigm for weakly-supervised laparoscopic image segmentation. Extensive evaluations have demonstrated our method’s potential solution to this task and its adaptability to different imaging modalities.

A primary limitation of our method is high computational demand. Future efforts will aim to lower computational expenses. Moreover, we simulated the sparse annotations due to the lack of real weak labels, inspired by previous works [3,20,23,5]. Thus, applying our method to more datasets with real weak labels, which more closely mirror real-world scenarios, is an aspect of future work. Additionally, we urge both ourselves and the community to contribute datasets featuring real weak labels to facilitate continued studies in this area.

**Acknowledgments.** This work was supported by JSPS KAKENHI (24H00720, 24K03262), JST CREST (JPMJCR20D5), JST [Moonshot R&D] (JPMJMS2033, JPMJMS2214), and JSPS Bilateral Joint Research Project.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Bernard, O., Lalonde, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Gonzalez Ballester, M.A., Sanroma, G., Napel, S., Petersen, S., Tziritas, G., Grinias, E., Khened, M., Kollerathu, V.A., Krishnamurthi, G., Rohé, M.M., Pennec, X., Sermesant, M., Isensee, F., Jäger, P., Maier-Hein, K.H., Full, P.M., Wolf, I., Engelhardt, S., Baumgartner, C.F., Koch, L.M., Wolterink, J.M., Išgum, I., Jang, Y., Hong, Y., Patravali, J., Jain, S., Humbert, O., Jodoin, P.M.: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging* **37**(11), 2514–2525 (2018)
2. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**(9), 1124–1137 (2004)
3. Fuentes-Hurtado, F., Kadkhodamohammadi, A., Flouty, E., Barbarisi, S., Luengo, I., Stoyanov, D.: Easylabels: weak labels for scene segmentation in laparoscopic videos. *International journal of computer assisted radiology and surgery* **14**, 1247–1257 (2019)
4. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059 (2016)

5. Gao, F., Hu, M., Zhong, M.E., Feng, S., Tian, X., Meng, X., yi-di-li Ni-jia ti, M., Huang, Z., Lv, M., Song, T., Zhang, X., Zou, X., Wu, X.: Segmentation only uses sparse annotations: Unified weakly and semi-supervised learning in medical images. *Medical Image Analysis* **80**, 102515 (2022)
6. Grady, L.: Random walks for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**(11), 1768–1783 (2006)
7. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. In: *Advances in neural information processing systems* (2004)
8. Hong, W.Y., Kao, C.L., Kuo, Y.H., Wang, J.R., Chang, W.L., Shih, C.S.: Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. arXiv preprint arXiv:2012.12453 (2020)
9. Kim, B., Ye, J.C.: Mumford–shah loss functional for image segmentation with deep learning. *IEEE Transactions on Image Processing* **29**, 1856–1866 (2019)
10. Lafferty, J.D., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: *Proceedings of the Eighteenth International Conference on Machine Learning*. pp. 282–289 (2001)
11. Lee, H., Jeong, W.K.: Scribble2label: Scribble-supervised cell segmentation via self-generating pseudo-labels with consistency. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention, LNCS*, vol 12261. pp. 14–23 (2020)
12. Liu, X., Yuan, Q., Gao, Y., He, K., Wang, S., Tang, X., Tang, J., Shen, D.: Weakly supervised segmentation of COVID19 infection with scribble annotation on CT images. *Pattern Recognition* **122**, 108341 (2022)
13. Luo, X., Hu, M., Liao, W., Zhai, S., Song, T., Wang, G., Zhang, S.: Scribble-supervised medical image segmentation via dual-branch network and dynamically mixed pseudo labels supervision. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention, LNCS*, vol 13431 (2022)
14. Marin, D., Tang, M., Ayed, I.B., Boykov, Y.: Beyond gradient descent for regularized segmentation losses. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 10187–10196 (2019)
15. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention, LNCS*, vol 9351. pp. 234–241 (2015)
16. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. In: *Advances in Neural Information Processing Systems*. vol. 28. Curran Associates, Inc. (2015)
17. Tang, M., Djelouah, A., Perazzi, F., Boykov, Y., Schroers, C.: Normalized cut loss for weakly-supervised CNN segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 1818–1827 (2018)
18. Tang, M., Perazzi, F., Djelouah, A., Ben Ayed, I., Schroers, C., Boykov, Y.: On regularized losses for weakly-supervised CNN segmentation. In: *Proceedings of the European Conference on Computer Vision*. pp. 507–522 (2018)
19. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging* **36**(1), 86–97 (2016)
20. Valvano, G., Leo, A., Tsiftaris, S.A.: Learning to segment from scribbles using multi-scale adversarial attention gates. *IEEE Transactions on Medical Imaging* **40**(8), 1990–2001 (2021)
21. Wang, J., Lukasiewicz, T.: Rethinking bayesian deep learning methods for semi-supervised volumetric medical image segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022)

22. Wang, Z., Lu, B., Long, Y., Zhong, F., Cheung, T.H., Dou, Q., Liu, Y.: Autolapar: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, LNCS, vol 13437. pp. 486–496 (2022)
23. Yang, Z., Simon, R., Linte, C.: A weakly supervised learning approach for surgical instrument segmentation from laparoscopic video sequences. In: Medical Imaging 2022: Image-Guided Procedures, Robotic Interventions, and Modeling. vol. 12034, p. 120341U (2022)
24. Zhang, T.Y., Suen, C.Y.: A fast parallel algorithm for thinning digital patterns. *Communications of the ACM* **27**(3), 236–239 (1984)