Anatomically-aware Uncertainty for Semi-supervised Image Segmentation

Sukesh Adiga V[a,*], Jose Dolz[a], Herve Lombaert[a]

[a]*ETS Montreal, Canada*

ARTICLE INFO

Article history:

Anatomically-aware Representation; Plausible Segmentation; Uncertainty Estimation; Self--ensembling; Semi-supervised Learning.

ABSTRACT

Semi-supervised learning relaxes the need of large pixel-wise labeled datasets for image segmentation by leveraging unlabeled data. A prominent way to exploit unlabeled data is to regularize model predictions. Since the predictions of unlabeled data can be unreliable, uncertainty-aware schemes are typically employed to gradually learn from meaningful and reliable predictions. Uncertainty estimation methods, however, rely on multiple inferences from the model predictions that must be computed for each training step, which is computationally expensive. Moreover, these uncertainty maps capture pixel-wise disparities and do not consider global information. This work proposes a novel method to estimate segmentation uncertainty by leveraging global information from the segmentation masks. More precisely, an anatomically-aware representation is first learnt to model the available segmentation masks. The learnt representation thereupon maps the prediction of a new segmentation into an anatomically-plausible segmentation. The deviation from the plausible segmentation aids in estimating the underlying pixel-level uncertainty in order to further guide the segmentation network. The proposed method consequently estimates the uncertainty using a single inference from our representation, thereby reducing the total computation. We evaluate our method on two publicly available segmentation datasets of left atria in cardiac MRIs and of multiple organs in abdominal CTs. Our anatomically-aware method improves the segmentation accuracy over the state-of-the-art semi-supervised methods in terms of two commonly used evaluation metrics.

1. Introduction

Segmentation is a fundamental task in medical image analysis, where image pixels are associated with a target object, such as an organ, structure, or abnormal region. It is a vital pre-processing step in many clinical applications, notably in computer-assisted diagnosis, intervention assistance, treatment planning, and personalized medicine (Duncan and Ayache, 2000; Ayache and Duncan, 2016). Recent segmentation methods based on deep learning techniques are driving progress under the full-supervision regime, often outperforming traditional methods (Litjens et al., 2017). Such a regime, however, relies on a large amount of annotations, which is time-consuming. Delineating an image at a pixel-level is indeed challenging, especially in homogeneous or low-contrast regions, and requires prohibitive clinical expertise. The burden of image

*Corresponding author: Sukesh Adiga V, **Email:** *sukesh.adiga-vasudeva.1@ens.etsmtl.ca*. All authors are with Computer and Software Engineering Department, ETS Montreal, 1100 Notre Dame St. W., Montreal QC, H3C 1K3, Canada

annotation motivates new learning strategies with limited supervision (Cheplygina et al., 2019).

Semi-supervised learning is an emerging strategy that alleviates annotation scarcity by leveraging unlabeled data with a small set of labeled data. Current semi-supervised segmentation methods typically utilize unlabeled data either in the form of pseudo labels (Bai et al., 2017; Zheng et al., 2020) or in a regularization term (Nie et al., 2018; Cui et al., 2019; Peng et al., 2020). The former strategies augment the original labeled dataset with unlabeled data alongside its corresponding model predictions, commonly referred to as pseudo labels. Later techniques incorporate unlabeled data into the training process by constraining predictions with a regularizer term. Training these semi-supervised approaches typically involves a supervised loss associated with labeled data and an unsupervised loss associated with unlabeled data.

Among regularization techniques, consistency-based approaches (Laine and Aila, 2017; Tarvainen et al., 2017) are often used in semi-supervision due to simple ways to leverage unlabeled data. Their approach encourages two or more segmentation predictions to be consistent under different perturbations of the input data (Cui et al., 2019; Bortsova et al., 2019; Li et al., 2020b). However, the segmentation predictions can be unreliable and noisy for unlabeled data since its annotations are unavailable. To alleviate this issue, uncertainty-aware regularization methods (Yu et al., 2019; Sedai et al., 2019) have been proposed to gradually add reliable target regions in predictions. Although these methods perform well in low-labeled data regime, their high computation and complex training techniques remain a limiting factor to broader applications. For instance, the pixel-level uncertainty approximation with Monte-Carlo Dropout (MCDO) (Gal and Ghahramani, 2016) or ensembling (Lakshminarayanan et al., 2017) requires multiple predictions per image, thereby increasing the computation of each training step. Moreover, these approaches do not consider global information to estimate uncertainty. The resulting uncertainty maps capture pixel-wise disparity, most likely around boundaries (Kendall et al., 2017). However, high gradient regions near anatomical boundaries or inner content of anatomical structures should have a certain labeling mask. For instance, Fig. 1 shows uncertainty captured by MCDO mostly over boundaries, while regions with high gradients (red arrows) could indicate certain boundaries or anatomical details with certainty. Probable uncertainty may lie in areas of low image gradients. For instance, anatomical boundaries may be unclear due to imaging or even non-existent in case of an arbitrary cut from an annotator (green arrows), as illustrated in the pulmonary veins in Fig. 1. Existing methods could benefit from capturing informative uncertainty in images beyond highlighting high image gradients or all over boundaries.

The global information of the anatomical regions is one promising direction to provide cues about informative uncertainty in images. Our approach will, therefore, exploit and capture global anatomical information by leveraging available masks to approximate segmentation uncertainty. Our main idea is to learn an anatomically-aware representation from a training set of segmentation masks. The learnt representation maps incorrect model predictions onto an anatomically-plausible segmentations. The plausible segmentation is subsequently used to estimate the uncertainty maps and further guide training of the segmentation network. We hypothesize that the proposed uncertainty estimates are more robust and computationally less expensive than deriving them from a standard entropy variance-based method, which requires multiple inferences for each training step.

Our contributions. We propose a novel approach to estimate the uncertainty maps from an anatomically-aware representation of the segmentation masks, in order to guide the training of a semi-supervised segmentation model. More precisely, we innovate semi-supervised segmentation with uncertainty-based training by integrating a pretrained denoising autoencoder (DAE) into the training of our segmentation network to: (i) map the inaccurate model predictions to plausible segmentation masks and (ii) estimate new uncertainty maps that guide the training of our segmentation model. As we approximate the uncertainty based on the difference between predicted segmentation and its DAE reconstruction learned from the segmentation mask, it can better integrate anatomical information. In contrast to most uncertainty-based approaches, estimating the uncertainty map requires a single inference from the DAE model, thereby reducing computational complexity. Our method is extensively evaluated on two medical imaging datasets: the 2018 Atrial segmentation challenge dataset (Xiong et al., 2021) and the 2021 Abdominal organ segmentation dataset (Ma et al., 2022). Results demonstrate the superiority of our approach over the state-of-the-art methods in semi-supervised segmentation.

A preliminary version of this work has been published in MICCAI 2022 (Adiga Vasudeva et al., 2022). This work includes a comprehensive literature review, extensive experiments, and a thorough discussion. The additional contributions in this manuscript are summarized as follows: (i) an additional multi-class abdominal segmentation dataset is evaluated for all our experiments, including ablation studies; (ii) the impact of various design choices made in the anatomically-aware representation prior (DAE) module are studied; (iii) a qualitative comparative analysis of uncertainty for different methods and their computation time are provided; (iv) additional related baseline that use a Monte-Carlo Dropout-based uncertainty estimation is provided for comparison (Wang et al., 2020); (v) the introduction and motivation of our approach are significantly extended with illustrations of our uncertainty maps; (vi) our literature review is expanded with recent uncertainty-aware as well as anatomically-plausible segmentation methods.
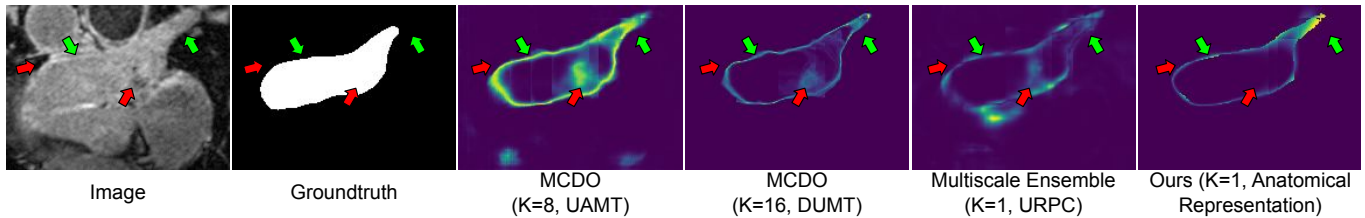
Figure 1: **Uncertainty maps from different semi-supervision methods**. K denotes the number of inferences. Green arrows in regions of probable uncertainty due to unclear boundaries or annotator cut preference (such as in pulmonary veins cut in top right). Red arrows in regions of lower uncertainty as they depict high image gradients in uninformative clear boundary or inner foreground content.

1.1. Related Work

Semi-Supervised Segmentation. Semi-supervised learning (SSL) is an established approach in the literature under the paradigm of learning with limited supervision (Jiao et al., 2022). A wide range of SSL strategies have been explored for segmentation, such as self-training (Bai et al., 2017; Zheng et al., 2020), entropy minimization (Grandvalet and Bengio, 2004; Wu et al., 2021), consistency regularization (Cui et al., 2019; Bortsova et al., 2019), co-training (Peng et al., 2020; Xia et al., 2020) or adversarial learning (Nie et al., 2018; Chaitanya et al., 2019). For instance, self-training methods (Bai et al., 2017; Zheng et al., 2020) typically employ pseudo-labels on unlabeled data to train models in an iterative way. However, potential labeling mistakes in the pseudo labels can quickly propagate during training, causing undesired segmentation outcomes. Entropy minimization strategies (Wu et al., 2021) circumvent such issues by enforcing a high confidence in predictions but can also easily lead to trivial solutions if additional priors are not used. Co-training approaches (Peng et al., 2020; Xia et al., 2020) avoid iterations but at the cost of simultaneously training two or more networks with multi-view images. Adversarial methods (Nie et al., 2018; Chaitanya et al., 2019) encourage the predictions of unlabeled images to be closer to those of the labeled images, however, they remain challenging in terms of convergence (Salimans et al., 2016). Among the existing SSL strategies, consistency regularization-based methods (Laine and Aila, 2017; Tarvainen et al., 2017) are popular due to their simple assumption that predictions should not change significantly under different realistic data perturbations. This notion is formulated as a consistency regularization term in the loss function, which encourages predictions to be consistent between data and its perturbed version (Cui et al., 2019; Bortsova et al., 2019; Li et al., 2020b). Similarly, our method leverages unlabeled data with a consistency regularizer.

Uncertainty-based methods. The uncertainty estimation approaches often employ Bayesian neural networks (Neal, 2012), however, their training process poses significant computational challenges. Recent deep learning methods address this limitation by approximating uncertainty through the generation of multiple samples (Abdar et al., 2021). For instance, Monte-Carlo Dropout (MCDO) (Gal and Ghahramani, 2016) performs several forward passes through the same model with dropout enabled at test time to generate multiple samples for the same input. Whereas a deep ensemble (Lakshminarayanan et al., 2017) trains a set of independent models to generate multiple samples. These approaches, however, tackle the problem of approximating *epistemic* uncertainty associated with the model output but not the *aleatoric* uncertainty associated with the model input (Kendall and Gal, 2017). A set of recent methods models the *aleatoric* uncertainty by using intra-/inter-annotation variability as a proxy to the underlying input uncertainties (Kohl et al., 2018; Baumgartner et al., 2019; Monteiro et al., 2020). All of the aforementioned methods have been shown to produce reliable uncertainty estimations in fully-supervised segmentation (Mehta et al., 2022; Camarasa et al., 2021).

In the context of semi-supervised segmentation, the uncertainty in the prediction is widely used within the optimization process (Yu et al., 2019; Wang et al., 2020, 2021). In particular, the uncertainty information assists the segmentation models by providing reliable target regions on unlabeled data during each training step. For instance, Yu et al. (2019) first approximates an uncertainty map using a predictive entropy of several predictions under data and model perturbations. The generated uncertainty map is later used to gradually add the reliable target regions in the consistency loss term. This idea was further extended to integrate uncertainty on a feature-level (Wang et al., 2020) and multiple prediction branches (Wang et al., 2022). The uncertainty estimation in these approaches commonly use MCDO (Gal and Ghahramani, 2016) or ensembling (Lakshminarayanan et al., 2017), which inherently relies on multiple predictions per image. In addition to being computationally expensive, estimating such entropy-based uncertainty is suboptimal in a multi-class scenario since it disregards inter-class overlaps (Van Waerebeke et al., 2022). More recently, multi-scale (Luo et al., 2022) or multi-decoder (Wu et al., 2022) approaches have been proposed to overcome the expensive computation of uncertainty using multiple predictions in a single forward pass. Nevertheless, these methods failed to capture the actual uncertainty regions. In contrast to existing strategies, our method leverages an anatomically-aware representation from the available annotations to estimate the uncertainty in a single inference step. This strategy leads to a lower computational complexity and an improved computational efficiency.

Towards anatomically-plausible segmentations. Recent approaches incorporate anatomically-aware priors in a segmentation network (Oktay et al., 2017; Ravishankar et al., 2017; Painchaud et al., 2020) by learning the variability of structures in a medical imaging dataset. For instance, Oktay et al. (2017) first learn an anatomically-aware representation with an autoencoder-based architecture using segmentation masks. This representation is later utilized to map a prediction into an anatomically-plausible space. These methods use the encoder of the representation as a global shape regularizer that enforces the model predictions to follow the ground truth distribution. The anatomically-aware representation can also map an erroneous mask into an anatomically-plausible segmentation. Such mapping is subsequently used to correct the segmentation predictions as a post-processing step (Larrazabal et al., 2020; Painchaud et al., 2020) or improve the segmentation on unseen test images (Karani et al., 2021). In order to encode the masks in the anatomically-aware representation, a substantial amount of annotations are used either from the given dataset (Larrazabal et al., 2020; Painchaud et al., 2020) or the source domain dataset (Karani et al., 2021). The anatomically-aware representation is alternately substituted with a probabilistic atlas to enforce the priors (Zheng et al., 2019; Huang et al., 2022), which requires an aligned dataset. For instance, Dalca et al. (2018) learns an anatomically-aware representation on aligned labelings and subsequently uses it for unsupervised segmentation on aligned images. In contrast to these approaches, our method leverages an anatomically-aware representation in a low-data regime with the goal of obtaining uncertainty maps in order to guide the segmentation network during the training process.

2. Method

An overview of the proposed anatomically-aware uncertainty estimation for semi-supervised segmentation is shown in Fig 2. The main idea is to exploit an anatomically-aware representation that maps the segmentation prediction into a plausible mask. The reconstructed segmentation will be indicative in estimating an uncertainty map, which later is used to guide the segmentation training. The following subsections describe the semi-supervised setting, anatomically-aware representation and uncertainty estimation process.

2.1. Preliminaries

The standard semi-supervised learning consists of N labeled and M unlabeled data in the training set, where $N \ll M$. Let $D_L = \{(x_i, y_i)\}_{i=1}^N$ and $D_U = \{(x_i)\}_{i=(N+1)}^{(N+M)}$ denote the labeled and unlabeled sets, where an input volume is represented as $x_i \in \mathbb{R}^{H \times W \times D}$ and its corresponding segmentation mask is $y_i \in \{0, 1, ..., C\}^{H \times W \times D}$, with C being the number of classes. The objective is to train

a segmentation network with a combination of supervised loss $\mathcal{L}_s$ and unsupervised loss $\mathcal{L}_u$ using labeled and unlabeled data, i.e., $\mathcal{L} = \mathcal{L}_s + \lambda \mathcal{L}_u$, where λ controls the weight of unsupervised loss.

2.2. Mean Teacher Formulation

Following current literature (Yu et al., 2019), we adopt the common mean teacher approach (Tarvainen et al., 2017) for training a segmentation network. It consists of a student (S) and a teacher (T) model, both having the same segmentation architecture. The overall objective function is defined as follows:

$$\mathcal{L} = \min_{\theta_S} \sum_{i=1}^N \mathcal{L}_s(f(x_i; \theta_S), y_i) + \\ \lambda_c \sum_{i=1}^{N+M} \mathcal{L}_c(f(x_i, \eta; \theta_S), f(x_i, \eta'; \theta_T)), \quad (1)$$

where $f(\cdot)$ denotes the segmentation network, and θ_S and θ_T are the learnable weights of the student and teacher models. The supervised loss $\mathcal{L}_s$ measures the segmentation quality on the labeled data, whereas the unsupervised consistency loss $(\mathcal{L}_c = \mathcal{L}_u)$ measures the prediction consistency between the student and the teacher models for the same input volume x_i under different perturbations (η and η'). The balance between the supervised and unsupervised loss is controlled by a ramp-up weighting coefficient λ_c, which is defined as

$$\lambda_c = \beta * e^{-r(1 - \frac{t}{t_{max}})^2}, \quad (2)$$

where β is a consistency weight, r controls the rate of ramp-up, t and t_{max} denote the current and maximum training steps. For training, the student model parameters (θ_S) are optimized with stochastic gradient descent (SGD), whereas the teacher model parameters (θ_T) are updated using an exponential moving average (EMA) at each training step t. The EMA is defined as

$$\theta_T^t = \alpha \theta_T^{t-1} + (1 - \alpha)\theta_S^t, \quad (3)$$

where α is the smoothing coefficient of EMA that controls the update rate.

2.3. Anatomically-aware Uncertainty Approach

The reliability of the model prediction on the unlabeled dataset plays an essential role in the consistency loss. An uncertainty-aware scheme can assist this loss by providing reliable target regions. The existing approaches (Yu et al., 2019; Wang et al., 2020) estimate uncertainty at a pixel-level, which fails to consider global information within the dataset. To address this limitation, our approach learns an anatomically-aware representation prior in order to capture global information. The measurable deviations from this prior provide informative cues about the uncertainty of the segmentation mask. The following subsections elaborate on our anatomically-aware uncertainty method.
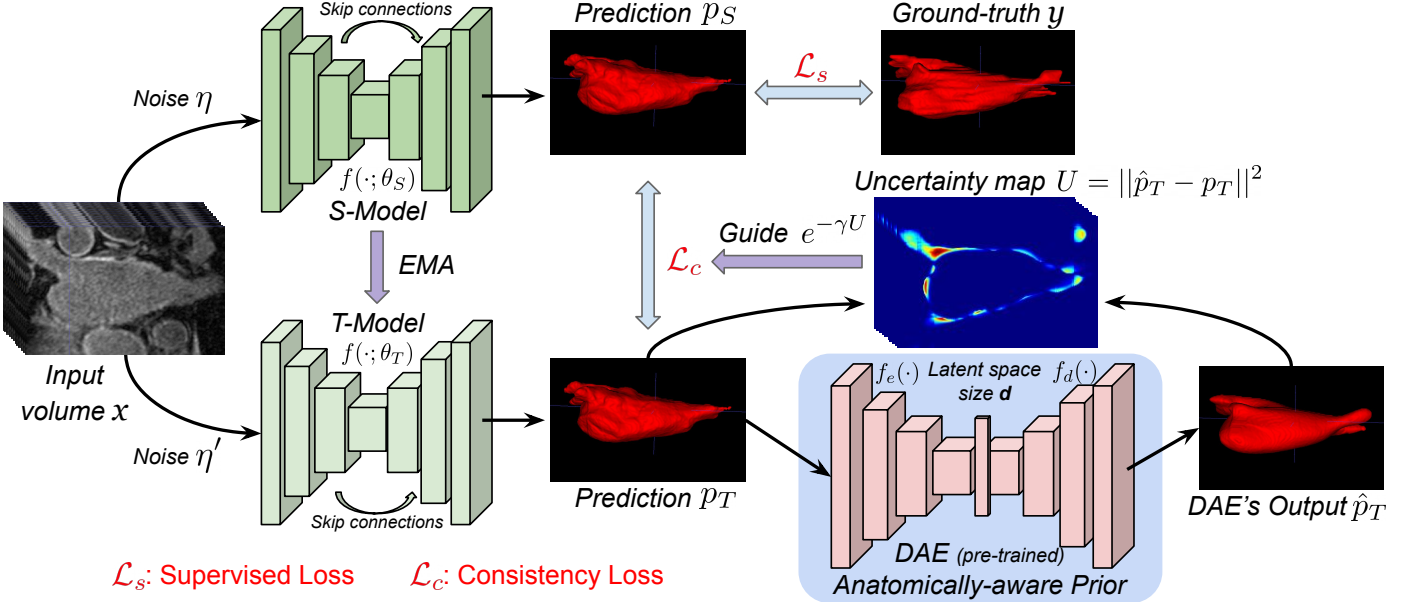
Figure 2: **Overview of our uncertainty estimation from anatomically-aware representation for semi-supervised segmentation.** A pre-trained anatomically-aware representation (i.e., a DAE) module is integrated into the training of the mean teacher model, which maps the teacher prediction p_T into a plausible segmentation $\hat{p}_T$. The uncertainty map (U) is subsequently estimated with the output of the teacher and the DAE model in order to further guide the student model. Best viewed in colour.

2.3.1. Anatomically-aware Representation Prior

Incorporating anatomically-aware prior in deep segmentation models is not obvious. One of the reasons is that, in order to integrate such prior knowledge during training, one needs to augment the learning objective with a differentiable term, which is not trivial. To circumvent these difficulties, a simpler solution is to resort to an autoencoder trained with segmentation masks, which maps the predictions into anatomically-plausible segmentation. This strategy has been adopted for fully-supervised learning as a global regularizer during training in (Oktay et al., 2017) and as a post-processing step in (Larrazabal et al., 2020) to correct the segmentation predictions. Motivated by this concept, we encode the available segmentation masks in a non-linear latent space of a denoising autoencoder (DAE) (Vincent et al., 2010) to learn an anatomically-aware representation prior. This learnt representation captures the global information from the segmentation masks such that it maps an inaccurate prediction into a plausible segmentation.

The DAE model consists of an encoder $f_e(\cdot)$ and a decoder $f_d(\cdot)$ with a d-dimensional latent space as shown in the Fig. 2. The DAE is trained to reconstruct the clean label y_i from its corrupted version $\tilde{y}_i$, which can be achieved with a mean squared error loss: $\frac{1}{H \times W \times D} \sum_v ||f_d(f_e(\tilde{y}_{i,v})) - y_{i,v}||^2$, where v is a voxel. Additionally, the dice loss is added to handle the class imbalance between foreground and background in the labels.

2.3.2. Anatomically-aware Uncertainty

The role of the uncertainty is to gradually update the student model with reliable target regions from the teacher

model predictions. Our proposed method estimates the uncertainty directly from the anatomically-aware representation network $f_d(f_e(\cdot))$, requiring only one inference step. First, we map the segmentation prediction from the teacher model (p_{T_i}) with a DAE model to produce a plausible segmentation $\hat{p}_{T_i} = f_d(f_e(p_{T_i}))$. We subsequently estimate the uncertainty as the pixel-wise difference between the DAE output and the prediction, which is given as:

$$U_i = ||\hat{p}_{T_i} - p_{T_i}||^2. \tag{4}$$

Note that the uncertainty formulation is related to the conventional sample variance-based uncertainty estimation. Specifically, for a given input, x_i, and its corresponding multiple model predictions, p_{i_s}, the sample variance estimation is defined as follows:

$$var(p_i) = \frac{1}{S-1} \sum_{s=1}^{S} (p_{i_s} - \bar{p}_i)^2,$$

where $\bar{p}_i$ represents the sample mean and is defined as $\bar{p}_i = \frac{1}{S} \sum_{s=1}^{S} (p_{i_s})$. The parameter S denotes the number of prediction samples. When S is set to 2, the sample mean $\bar{p}_i$ reduces to $\frac{p_{i_1} + p_{i_2}}{2}$, resulting in the variance estimation taking the form of:

$$var(p_i) = (p_{i_1} - \frac{p_{i_1} + p_{i_2}}{2})^2 + (p_{i_2} - \frac{p_{i_1} + p_{i_2}}{2})^2,$$
$$= (\frac{p_{i_1} - p_{i_2}}{2})^2 + (\frac{p_{i_2} - p_{i_1}}{2})^2,$$
$$var(p_i) = \frac{1}{2}(p_{i_1} - p_{i_2})^2.$$

The above equation is equivalent to our uncertainty formulation in Eq. 4, where two samples are drawn from the output of the teacher model and the DAE model.

The resulting uncertainty maps from Eq. 4 are subsequently used to obtain the reliable target regions as follows: $e^{-\gamma U_i}$, similarly to (Luo et al., 2022), where γ is an uncertainty weighting factor empirically set to 1. The reliable targets are finally combined in a consistency loss as:

$$\mathcal{L}_c(p_{S_i}, p_{T_i}) = \frac{\sum_v e^{-\gamma U_{i,v}} ||p_{S_{i,v}} - p_{T_{i,v}}||^2}{\sum_v e^{-\gamma U_{i,v}}}, \qquad (5)$$

where v is a voxel. Note that the consistency loss $\mathcal{L}_c$ will be equivalent to a standard mean teacher method (Tarvainen et al., 2017) when $\gamma = 0$. Overall, we jointly optimize the consistency loss $\mathcal{L}_c$ and supervised loss $\mathcal{L}_s$ as learning objectives, where $\mathcal{L}_s$ is a combination of cross-entropy and dice losses.

3. Experiments

3.1. Datasets

The performance of our method is validated on two publicly available benchmarks: (a) the left atrium (LA) binary segmentation dataset from the 2018 atrial challenge (Xiong et al., 2021), and (b) the abdominal multi-organ segmentation dataset from the FLARE challenge (Ma et al., 2022).

(a) LA dataset. It consists of 100 3D late gadolinium-enhanced magnetic resonance imaging (LGE-MRI) scans and corresponding LA segmentation masks. These scans have an isotropic resolution of 0.625 mm^3 and are center cropped at the heart region. The dataset is split into 80 for training and the remaining 20 for testing as in the literature (Yu et al., 2019; Li et al., 2020a; Wang et al., 2020; Luo et al., 2021).

(b) FLARE dataset. This dataset consists of 361 CT scans of the abdominal region and corresponding segmentation masks of four organs, namely liver, kidney, spleen, and pancreas. These scans are collected from multiple medical centers, having varying resolutions. Each image is first resampled to a uniform resolution of $2 \times 2 \times 2.5$ mm^3 and then normalized by clipping the intensity values outside $[0.5, 0.95]$ percentile range. For all our experiments, we use a fixed dataset split of 260 for training, 26 for validation, and the remaining 75 for testing.

3.2. Implementation and Training details

To validate our proposed method, we employ a V-Net (Milletari et al., 2016) as a backbone architecture for the segmentation networks, as followed in earlier work (Yu et al., 2019; Wang et al., 2020; Luo et al., 2021). Our anatomically-aware representation prior module (i.e., a DAE) follows a similar architecture as V-Net but without skip connections. Such design effectively makes it an autoencoder-style architecture, which is also comparable to prior work (Oktay et al., 2017; Larrazabal et al., 2020).

To encode the segmentation mask in a latent space, a dense layer of d-dimension is added at the bottleneck layer of the DAE module as shown in Fig. 2. For training, the student model uses a SGD optimizer with an initial learning rate (lr) of 0.1 and a momentum of 0.9 with a cosine annealing decaying (Loshchilov and Hutter, 2017). The teacher weights (in Eq. 3) are updated by an EMA with a rate of $\alpha = 0.99$ (Tarvainen et al., 2017). The DAE model is also trained using a SGD optimizer with an initial $lr = 0.1$, a momentum of 0.9, and decaying the lr by a factor of 2 every 5000 iterations. Following the literature (Yu et al., 2019; Luo et al., 2022), the consistency weight β and ramp-up factor r in Eq. 2 are set to 0.1 and 5, respectively. Inputs to both segmentation and DAE networks are randomly cropped to a size of $112 \times 112 \times 80$ and $144 \times 144 \times 96$ for LA and FLARE datasets, respectively. We employ online standard data augmentation techniques such as random flipping and rotation. In addition, input labels to the DAE are corrupted with a random swapping of pixels around class boundaries, morphological operations (erosion and dilation), resizing, and adding/removing basic shapes (Van der Walt et al., 2014). The latent space of the DAE is injected with a small noise drawn from a Gaussian distribution to explore different sets of plausible segmentation during training of the segmentation network. The training set is partitioned into N labeled and M unlabeled splits, which are fixed across all experiments. The batch size is set to 4 in both networks. Input batch for the segmentation network uses two labeled and unlabeled data. During the inference phase, the segmentation predictions are generated using the sliding window strategy. For the cardiac dataset (LA), following the literature (Yu et al., 2019; Li et al., 2020a; Luo et al., 2021), the final model is evaluated at the last training iteration (i.e., 6000), whereas the best validation model is selected in the case of the abdominal dataset (FLARE). All our experiments were run on an NVIDIA RTX A6000 GPU with PyTorch 1.8.0. The implementation of our work is available at: https://github.com/adigasu/Anatomically-aware_Uncertainty_for_Semi-supervised_Segmentation.

3.3. Evaluation Metrics

We employ common Dice Score Coefficient (DSC) and 95% Hausdorff Distance (HD) evaluation metrics to assess quantitative segmentation performance. The DSC score evaluates the degree of overlap between ground truth and prediction regions. In contrast, the HD score measures the distance between ground truth and predicted segmentation boundaries. For a fair comparison, all experiments are run three times with a fixed set of seeds on the same machine, and their average results are reported.

Table 1: **Segmentation results on the LA test set for the 10% and 20% annotation settings**. Uncertainty-based methods with K inferences per training step are grouped at the bottom of each section, while $K = $ - indicates non-uncertainty-based methods. Ours achieves the best Dice (DSC) and Hausdorff (HD) scores in both annotation scenarios. The best and second-best results are highlighted in bold and underlined, whereas the statistical significance between the top two results is denoted in *. The number of labeled and unlabeled data indicated with N and M, respectively.

N/M	Methods	#K	DSC (%) ↑	HD (mm) ↓
80/0	Upper bound	-	91.23 ± 0.44	6.08 ± 1.84
8/0	Lower bound	-	76.07 ± 5.02	28.75 ± 0.72
	MT (Tarvainen et al., 2017)	-	78.22 ± 6.89	16.74 ± 4.80
	SASSnet (Li et al., 2020a)	-	83.70 ± 1.48	16.90 ± 1.35
8/72	DTC (Luo et al., 2021)	-	83.10 ± 0.26	12.62 ± 1.44
(10%)	UAMT (Yu et al., 2019)	8	$\underline{85.09 \pm 1.42}$	18.34 ± 2.80
	DUMT (Wang et al., 2020)	16	82.97 ± 1.76	14.43 ± 0.67
	URPC (Luo et al., 2022)	1	84.47 ± 0.31	17.11 ± 0.60
	Ours	1	$\mathbf{86.58 \pm 1.03^*}$	$\mathbf{11.82 \pm 1.42}$
16/0	Lower bound	-	81.46 ± 2.96	23.61 ± 4.94
	MT (Tarvainen et al., 2017)	-	86.06 ± 0.81	11.63 ± 3.40
	SASSnet (Li et al., 2020a)	-	87.81 ± 1.45	$\underline{10.18 \pm 0.55}$
16/64	DTC (Luo et al., 2021)	-	87.35 ± 1.26	10.25 ± 2.49
(20%)	UAMT (Yu et al., 2019)	8	87.78 ± 1.03	11.10 ± 1.91
	DUMT (Wang et al., 2020)	16	87.42 ± 0.97	10.78 ± 2.26
	URPC (Luo et al., 2022)	1	$\underline{88.58 \pm 0.10}$	13.10 ± 0.60
	Ours	1	$\mathbf{88.60 \pm 0.82}$	$\mathbf{7.61 \pm 0.78^*}$

4. Results

4.1. Comparison with the state-of-the-art

We first compare our method with relevant semi-supervised segmentation approaches and report the quantitative results in Tables 1 and 2. The upper and lower bound from the backbone architecture V-Net (Milletari et al., 2016) are reported at the top of each section. Furthermore, non-uncertainty-based methods such as MT (Tarvainen et al., 2017), DTC (Luo et al., 2021), and SASSnet (Li et al., 2020a) and uncertainty-based methods UAMT (Yu et al., 2019), DUMT (Wang et al., 2020), and URPC (Luo et al., 2022) are included in our evaluation.

(a) Left Atrium segmentation. Table 1 shows the segmentation performance on the Left Atrium (LA) test set under the standard 10% (top) and 20% (bottom) annotation settings. From the top half of the table, we observe that leveraging unlabeled data improves the lower bound across all models. The uncertainty-based approaches typically outperform their non-uncertainty counterparts in terms of DSC, but yield inferior results in terms of HD. Among these methods, UAMT and DTC achieve the best DSC and HD scores, respectively. Nevertheless, compared to these best-performing baselines, our method brings improvements in both DSC (1.5%) and HD (0.8mm) scores. Moreover, uncertainty estimation in our method requires a single inference from an anatomically-aware representation, whereas UAMT uses K=8 inferences per training step to obtain an uncertainty map. This highlights the efficiency of the proposed approach, which yields a better segmentation performance yet requires substantially less computational time at each training step.

Furthermore, we validate our method on the 20% annotation scenario, whose results are reported in bottom half of Table 1. We observe a similar trend in these results, with uncertainty-based approaches outperforming non-uncertainty-based methods in DSC, whereas their performance in terms of HD is degraded. An interesting observation is that existing methods are ranked differently across the two annotation settings, indicating that they might be sensitive to the annotation scenario. For example, while UAMT achieves the best DSC score under the 10% annotation setting, URPC yields the best results in the 20% annotation case. Similarly, the best models are different for HD metric, i.e., DTC under the 10% setting and SASSNet in the 20% setting. In contrast, our method consistently outperforms each existing approach in both DSC and HD scores, highlighting its robustness against the amount of labeled data.

(b) Abdominal multi-organ segmentations. Table 2 presents the performance of the abdominal multi-organ segmentations on the FLARE test set. The results of 10% and 20% annotation experiments are grouped in the top and bottom half of the table, respectively. We report individual organs as well as average results. From the top half of the table, we first notice that the performance of most existing methods is improved when compared to the lower bound in both DSC and HD scores, except SASSNet, DTC, and DUMT. The gap in the segmentation performance of SASSNet and DTC is due to the use of signed distance maps (SDM), which are designed for binary segmentation. Adopting these methods for multi-class segmentation is challenging since it requires careful hyperparameter tuning of per-class SDM predictions, which is beyond the scope of this work. Note that DUMT did not outperform the simple baseline under a multi-class setting, which is consistent with the observations in (Van Waerebeke et al., 2022). Among the existing methods, the uncertainty-based methods

Table 2: **Segmentation results on the FLARE test set for the 10% and 20% annotation settings.** Uncertainty-based methods with K inferences per training step are grouped at the bottom of each section, while $K = $ - indicates non-uncertainty-based methods. Our method produces the best results on average. The best and second-best results are highlighted in bold and underlined, whereas the statistical significance between the top two results is denoted in *. The number of labeled and unlabeled data indicated with N and M, respectively.

	N/M	Methods	#K	Average	Liver	Kidney	Spleen	Pancreas
DSC (%) ←	260/0	Upper bound	-	85.80 ± 1.42	94.95 ± 0.30	93.20 ± 0.81	89.65 ± 2.91	65.38 ± 2.57
	26/0	Lower bound	-	70.09 ± 2.77	88.37 ± 2.31	81.12 ± 2.49	70.74 ± 4.41	40.14 ± 3.84
		MT (Tarvainen et al., 2017)	-	70.76 ± 2.79	88.77 ± 3.11	83.34 ± 1.22	72.91 ± 4.35	38.01 ± 2.62
		SASSnet (Li et al., 2020a)	-	61.43 ± 14.3	86.94 ± 2.88	63.59 ± 43.0	59.83 ± 18.6	35.36 ± 5.05
	26/234 (10%)	DTC (Luo et al., 2021)	-	68.07 ± 1.42	87.99 ± 1.79	83.11 ± 3.93	66.04 ± 3.40	35.15 ± 1.26
		UAMT (Yu et al., 2019)	8	73.63 ± 0.65	**91.65 ± 0.49**	84.70 ± 2.39	76.16 ± 2.58	42.01 ± 2.24
		DUMT (Wang et al., 2020)	16	69.04 ± 1.39	87.28 ± 0.82	80.47 ± 3.88	68.23 ± 6.79	40.18 ± 2.59
		URPC (Luo et al., 2022)	1	73.31 ± 1.11	91.09 ± 0.62	85.88 ± 1.82	75.40 ± 2.64	40.89 ± 4.05
		Ours	1	**75.28 ± 1.54***	90.78 ± 1.26	**87.09 ± 1.89**	**78.13 ± 1.23**	**45.12 ± 2.20***
HD (mm) →	260/0	Upper bound	-	6.37 ± 1.15	5.50 ± 2.86	3.31 ± 1.10	7.49 ± 1.94	9.17 ± 0.66
	26/0	Lower bound	-	18.51 ± 4.01	15.26 ± 0.90	9.89 ± 2.13	30.51 ± 11.9	18.40 ± 3.53
		MT (Tarvainen et al., 2017)	-	18.58 ± 1.66	12.09 ± 3.72	8.70 ± 0.85	35.89 ± 7.47	17.64 ± 1.53
		SASSnet (Li et al., 2020a)	-	27.76 ± 8.51	24.59 ± 23.0	15.1 ± 11.1	51.86 ± 21.3	19.53 ± 0.89
	26/234 (10%)	DTC (Luo et al., 2021)	-	23.11 ± 6.01	21.63 ± 16.7	18.8 ± 11.3	32.64 ± 16.8	19.31 ± 2.07
		UAMT (Yu et al., 2019)	8	14.30 ± 1.94	**10.44 ± 1.45**	8.08 ± 1.41	20.44 ± 6.18	18.24 ± 3.04
		DUMT (Wang et al., 2020)	16	22.35 ± 3.82	13.23 ± 2.28	19.21 ± 13.9	36.17 ± 15.5	20.77 ± 3.58
		URPC (Luo et al., 2022)	1	14.23 ± 1.97	11.71 ± 2.37	**7.41 ± 1.16**	20.82 ± 5.02	16.96 ± 3.00
		Ours	1	**13.69 ± 0.68**	10.85 ± 1.69	9.48 ± 2.10	**18.45 ± 4.17**	**15.98 ± 1.33**
DSC (%) ←	52/0	Lower bound	-	70.15 ± 1.58	88.40 ± 1.24	81.91 ± 2.07	68.40 ± 5.68	41.88 ± 7.44
		MT (Tarvainen et al., 2017)	-	72.10 ± 1.84	89.82 ± 2.30	85.15 ± 1.66	71.87 ± 4.28	41.55 ± 2.99
		SASSnet (Li et al., 2020a)	-	69.74 ± 4.43	88.41 ± 1.10	86.19 ± 3.13	64.11 ± 12.1	40.25 ± 3.07
	52/208 (20%)	DTC (Luo et al., 2021)	-	68.49 ± 1.30	89.61 ± 0.71	83.31 ± 4.39	62.76 ± 5.64	38.29 ± 3.38
		UAMT (Yu et al., 2019)	8	74.72 ± 1.15	89.54 ± 3.10	87.92 ± 1.52	73.07 ± 3.91	**48.34 ± 1.41**
		DUMT (Wang et al., 2020)	16	72.08 ± 2.77	90.11 ± 1.66	85.43 ± 4.82	71.83 ± 0.92	40.94 ± 4.17
		URPC (Luo et al., 2022)	1	74.26 ± 1.02	91.02 ± 0.54	87.91 ± 2.47	72.06 ± 1.82	46.03 ± 0.40
		Ours	1	**76.69 ± 0.81***	**91.84 ± 1.00***	**88.72 ± 0.74**	**78.07 ± 0.69***	48.14 ± 1.73
HD (mm) →	52/0	Lower bound	-	15.63 ± 0.33	15.18 ± 4.46	11.93 ± 4.64	20.50 ± 2.56	**14.91 ± 2.78**
		MT (Tarvainen et al., 2017)	-	16.39 ± 3.34	**11.04 ± 0.58**	10.89 ± 0.91	25.70 ± 9.08	17.94 ± 4.50
		SASSnet (Li et al., 2020a)	-	23.84 ± 0.79	34.01 ± 14.3	11.89 ± 8.66	32.28 ± 1.53	17.16 ± 1.69
	52/208 (20%)	DTC (Luo et al., 2021)	-	22.46 ± 2.12	25.23 ± 20.1	18.09 ± 8.14	29.05 ± 4.84	17.46 ± 1.02
		UAMT (Yu et al., 2019)	8	14.50 ± 2.46	16.60 ± 4.11	7.83 ± 0.76	17.91 ± 8.34	15.66 ± 0.76
		DUMT (Wang et al., 2020)	16	15.53 ± 2.75	11.74 ± 2.27	8.64 ± 0.95	25.43 ± 8.42	16.31 ± 0.89
		URPC (Luo et al., 2022)	1	14.16 ± 0.68	11.16 ± 2.09	8.47 ± 2.79	20.66 ± 0.80	16.33 ± 1.70
		Ours	1	**13.11 ± 0.45**	11.32 ± 2.29	**7.79 ± 2.69**	**17.38 ± 4.19**	15.94 ± 0.28

(UAMT and URPC) perform well in both segmentation metrics. These methods improve the segmentation of liver and spleen regions, achieving the best average DSC and HD scores in UAMT and URPC, respectively. Compared to these best-performing baselines, our method predominantly improves the segmentation of challenging regions, notably the pancreas organ. Overall our anatomically-aware method consistently performs well in all regions and improves average DSC (1.65%) and HD (0.6mm) scores.

The results of the 20% annotation scenario are reported in the bottom half of Table 2. We notice a similar trend in the results when compared to the 10% annotation setting. All existing methods, except SASSNet and DTC, improve the segmentation performance over the lower bound in both DSC and HD scores. Our method outperforms the best-performing baselines (UAMT and URPC) in most cases and improves the average DSC (1.95%) and average HD (1mm) scores. These results show that our method consistently outperforms the existing approaches across different datasets and labeling scenarios. We can, therefore, argue that including our novel anatomically-

aware module is a valuable alternative to existing semi-supervised segmentation approaches.

4.2. Qualitative Analysis

Visual results of the left atrium (LA) segmentation obtained by different methods are depicted in Fig. 3. In the top row (10% annotation setting), the existing approaches produce segmentation output with holes (SASSnet, UAMT) and noisy boundaries (SASSnet, DTC, UAMT, DUMT). In contrast, URPC and our methods produce smoother segmentations, but URPC generates under-segmented output compared to our method. Note that a post-processing tool is commonly employed in SASSNet to improve the segmentation performance. However, this is avoided in our experiments for a fair comparison. In the 20% annotation setting (bottom row), with access to more labeled data, all methods reduce segmentation errors. Even in this case, our method produces promising and smoother segmentations when compared to existing approaches.

To highlight the deficiencies of these approaches in multi-class segmentation, we now show qualitative results
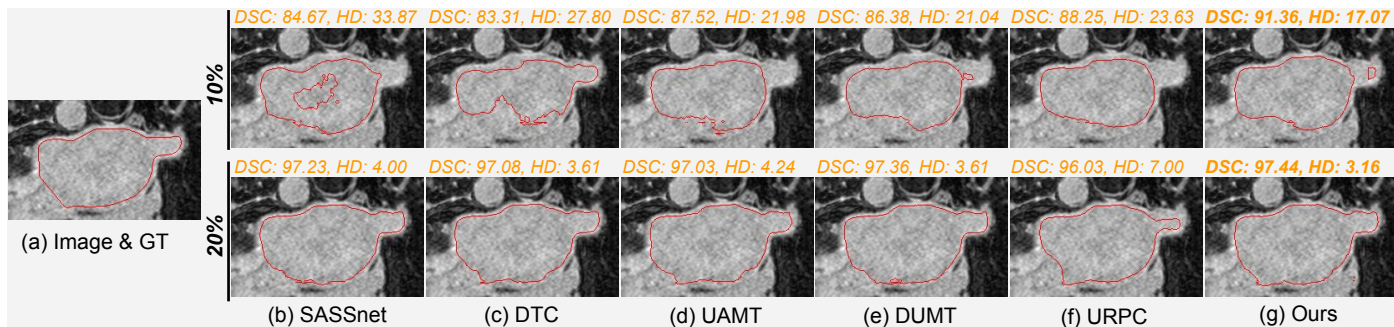
Figure 3: **Qualitative comparison under the 10% and 20% annotation settings on LA dataset.** DSC (%) and HD (mm) scores are mentioned at the top of each image. Each image is overlaid with a contour of segmentation prediction or ground truth (red).
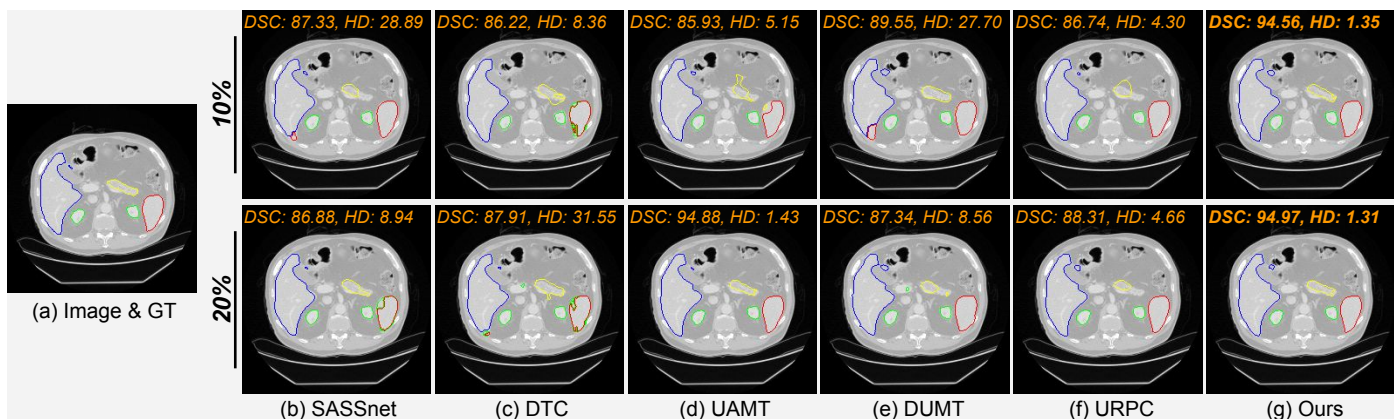

Figure 4: **Qualitative comparison under the 10% and 20% annotation settings on FLARE dataset.** Average DSC (%) and average HD (mm) scores are mentioned at the top of each image. The colorings are liver (blue), kidney (green), spleen (red), and pancreas (yellow).

on abdominal organs in Fig. 4. In the 10% annotation setting (top row), we first observe that misclassification between different organs is a common problem across existing approaches, notably in SASSnet, DTC, UAMT, and DUMT. For instance, part of the liver is segmented as a spleen in SASSnet and DUMT, whereas the parts of the spleen are misclassified as kidneys in DTC and as pancreas in UAMT. This misclassification could be due to either similar intensity characteristics across different organs (Durieux et al., 2018) or the inefficiency of networks in discriminating multi-class distributions (Van Waerebeke et al., 2022). Furthermore, most methods (SASSnet, DTC, UAMT, URPC) have failed to capture the challenging pancreas region. In contrast, our method provides an improved segmentation in this challenging region and minimizes classification errors. In the bottom row of Fig. 4, adding more labeled images to the training (20% annotation setting) also reduces classification errors (UAMT, URPC). Our method similarly improves the segmentation performance in all observed regions. The quantitative results from the previous section further support the superiority of our approach. Overall, we argue that the observed improvements in both datasets could be attributed to the knowledge derived from the anatomically-aware representation.
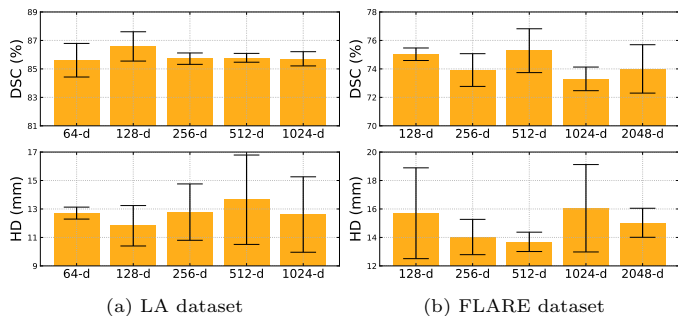

Figure 5: **Segmentation performance with different latent space sizes of DAE** - Each bar indicates the DSC (top) and HD (bottom) scores under the 10% annotation setting. The best results are obtained for the latent space size $d=128$ in binary LA segmentations (a), whereas $d=512$ is needed for abdominal multi-organ segmentations (b).

4.3. Choice of Latent Space in DAE

Our anatomically-aware prior (DAE) plays a vital role in guiding the segmentation model. Therefore, we investigate the impact of the design choices made in the DAE on the final segmentation performance. The latent space (LS) of our DAE is first studied under varying sizes (d) across two datasets in Fig. 5. The results show that the segmentation performance varies with LS sizes. The best results are achieved for $d=128$ in binary left atrium segmentations and $d=512$ in abdominal multi-organ segmentations. It
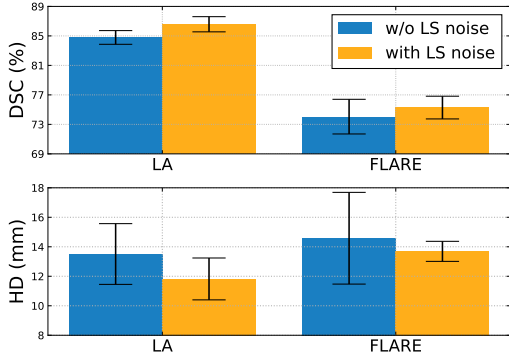
Figure 6: **Impact of noise in the latent space of DAE on segmentation performance** - Each bar indicates the DSC (top) and HD (bottom) scores under the 10% annotation setting. Addition of a noise (orange) in latent space improves DSC and HD scores.

indicates that the choice of the latent space size, d, depends on the complexity of the dataset.

Furthermore, the LS of the DAE is perturbed with an addition of a Gaussian noise. This facilitates a different set of reconstructions from the DAE when training the segmentation model. The different reconstructions aid in better guiding the segmentation model. To validate this notion, we conduct experiments with and without adding a noise in the LS across both datasets in Fig.6. The results demonstrate that the final segmentation performance improves up to 1.79% in DSC and 1.69mm in HD by adding a noise in the LS of the DAE module. These analyses show the impact of our design choices in the anatomically-aware prior on the segmentation performance.

4.4. Ablation Study on uncertainty

To validate the effectiveness of our uncertainty estimation on the segmentation performance, we conducted two experiments by adopting a threshold strategy and a predictive entropy scheme used in UAMT. Specifically, a threshold strategy filters out the most unreliable region from the uncertainty map (U_i), defined as $H > U_i$ with a threshold, H, set with a ramp-up function, as in UAMT (Yu et al., 2019). In the entropy experiments, we estimate the uncertainty (U_i) using the entropy of the DAE prediction ($\hat{p}_{T_i}$) and then combining it in a consistency loss as in Eq 5. The results of these ablation experiments on the LA and FLARE datasets under the 10% and 20% annotation settings are reported in Table 3. Compared to UAMT, our threshold and entropy experiments improve the segmentation performance in both DSC and HD scores in most cases. At the same time, our proposed uncertainty method (Sec. 2.3.2) achieves the best performance in all the settings. These results show the merit of our anatomically-aware uncertainty estimation for guiding the segmentation model.

4.5. Impact of γ and β hyperparameters

The sensitivity of the uncertainty weight γ (in Eq.5) and the consistency weight β on the segmentation performance

is shown in Fig. 7. In particular, we evaluate the segmentation performance using DSC and HD scores by varying the γ and β values across the LA and FLARE datasets. In Fig. 7(a)-(b), increasing the gamma value leads to an improvement in the segmentation performance in both DSC and HD scores across both datasets. The best results are usually observed for $\gamma = 1$. Beyond that, performance generally decreases, possibly due to an exponential decrease in the weight (Eq.5) of the reliable target regions.

Figure 7(c)-(d) shows the segmentation performance for varying the β values. The results show that increasing the beta value improves the segmentation performance. The best result is achieved for β=0.1 except in the LA dataset (in Fig. 7(c)), where β=1 produces the best scores. Nevertheless, we chose to set β=0.1 across all our experimental scenarios, as this value is widely adopted in the literature on consistency-based approaches (Tarvainen et al., 2017; Wang et al., 2021) and for a fair comparison with our baselines (Yu et al., 2019; Wang et al., 2020; Luo et al., 2022).

4.6. Training time

To evaluate the speed of our uncertainty estimation, we compare the computation time required for each training iteration by the proposed and the baseline methods in Table 4. From the table, we observe that the non-uncertainty-based methods (SASSnet, DTC) are slower when compared to uncertainty-based methods across both datasets, LA and FLARE. The relative slow speed of SASSnet and DTC is attributed to the additional computational overhead required for predicting the signed distance maps (SASSnet, DTC) and the inclusion of a discriminator module (SASSnet). On the other hand, ours and the URPC method are faster than the MCDO-based methods (UAMT and DUMT) due to the need of only one inference when estimating the uncertainty ($\#K$=1). Overall, our approach adds a minimal overhead on top of the mean teacher (MT) approach for estimating uncertainty while producing superior segmentation results on both datasets.

4.7. Uncertainty Analysis

The predicted segmentation and uncertainty map from different uncertainty-based methods are shown in Fig. 8. The top row shows the 10% annotation setting, where uncertainties are all over the predicted regions for UAMT. These uncertainties inside the prediction regions are reduced in DUMT, possibly due to more inferences and the addition of feature uncertainty. However, the uncertainties are highly focused on the prediction boundaries. The uncertainty is produced at arbitrary regions in URPC due to their multi-scale discrepancy-based uncertainty estimation. Our method produces uncertainty in challenging regions, such as unclear anatomical boundaries or annotator cuts (as in pulmonary veins), which are estimated using anatomically-aware representation. In the below row of

Table 3: **Effectiveness of our proposed uncertainty estimation on segmentation results using different strategies.** The number of labeled and unlabeled data indicated with N and M, respectively.

N/M	Methods	LA Dataset		FLARE Dataset	
		DSC (%) ↑	HD (mm) ↓	DSC (%) ↑	HD (mm) ↓
8/72 (10%)	UAMT (Yu et al., 2019)	85.09 ± 1.42	18.34 ± 2.80	73.63 ± 0.65	14.30 ± 1.94
	Ours (Threshold)	85.39 ± 0.91	12.96 ± 3.05	74.25 ± 1.76	14.47 ± 1.63
	Ours (Entropy)	85.92 ± 1.52	$\mathbf{11.16 \pm 0.82}$	74.01 ± 0.62	15.03 ± 2.00
	Ours	$\mathbf{86.58 \pm 1.03}$	11.82 ± 1.42	$\mathbf{75.28 \pm 1.54}$	$\mathbf{13.69 \pm 0.68}$
16/64 (20%)	UAMT (Yu et al., 2019)	87.78 ± 1.03	11.10 ± 1.91	74.72 ± 1.15	14.50 ± 2.46
	Ours (Threshold)	88.12 ± 1.16	8.44 ± 1.96	74.80 ± 0.80	14.09 ± 1.83
	Ours (Entropy)	87.76 ± 0.36	8.90 ± 0.48	74.57 ± 0.53	15.38 ± 2.57
	Ours	$\mathbf{88.60 \pm 0.82}$	$\mathbf{7.61 \pm 0.78}$	$\mathbf{76.69 \pm 0.81}$	$\mathbf{13.11 \pm 0.45}$

(a) LA - γ (b) FLARE - γ (c) LA - β (d) FLARE - β
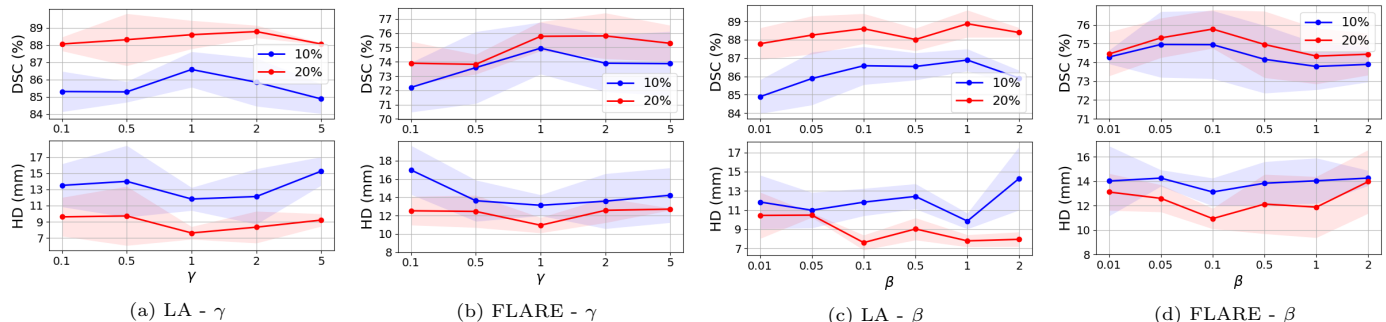
Figure 7: **Sensitivity of the consistency weight β (a, b) and the uncertainty weight γ (c, d)** - Each point in a line indicates the DSC (top) and HD (bottom) scores on LA and FLARE datasets under 10% (blue) and 20% (red) annotation settings.

Table 4: **Comparison of average training times in seconds per iteration.** Our method adds a minimal overhead on top of the mean teacher (MT) approach for uncertainty estimation.

Methods	#K	LA	FLARE
MT (Tarvainen et al., 2017)	-	0.612	1.108
SASSnet (Li et al., 2020a)	-	1.442	5.856
DTC (Luo et al., 2021)	-	0.989	4.874
UAMT (Yu et al., 2019)	8	1.207	2.429
DUMT (Wang et al., 2020)	16	3.804	7.678
URPC (Luo et al., 2022)	1	0.779	1.504
Ours	1	0.745	1.266

Fig. 8, increasing labeled samples (i.e., 20% setting) improves the predictions and uncertainty in most cases. Nevertheless, uncertainties are all over the boundaries, or arbitrary regions remain in the existing methods. Our method further improves the uncertainties due to the improvement of anatomically-aware representation using more access to labels. Moreover, our method requires single inference when compared to entropy-based methods.

5. Discussion and Conclusion

This work proposes a novel anatomically-aware uncertainty estimation method for semi-supervised image segmentation. Our approach consists of leveraging an anatomically-aware representation of labeling masks to estimate the segmentation uncertainty. The obtained uncertainty maps guide the training of the segmentation model within reliable regions of the predicted masks. Our experimental results demonstrate that the proposed method yields improved segmentation results when compared to state-of-the-art baselines on two publicly available benchmarks using left atria and abdominal organs. The qualitative results also show how our anatomically-aware approach improves segmentation in challenging image areas. The ablation studies demonstrate the effectiveness and robustness of our uncertainty estimation when compared to entropy-based methods. Adding noise in the latent space of our representation helps to map the predictions into a better set of plausible segmentations, which improves the segmentation accuracy. Unlike most uncertainty-based approaches, our anatomically-aware uncertainty requires a single inference, thereby reducing computational complexity. Moreover, as our anatomically-aware representation is independent of any image information, it can be further enhanced with existing segmentation masks from different datasets or imaging modalities (Karani et al., 2021), potentially further improving the modeling capacity of our representation. The learning representation with an additional constraint can also be explored separately as a post-processing tool that maps the erroneous prediction into anatomically-plausible segmentation (Larrazabal et al., 2020; Painchaud et al., 2020). Additionally, our anatomically-aware representation prior could also benefit from the image intensity information to learn a joint representation (Oktay et al., 2017; Judge et al., 2022) for uncertainty estimation in a limited supervision problem. Overall, our proposed approach could be leveraged to a broader range of applications where uncertainties could be related to anatomical information.
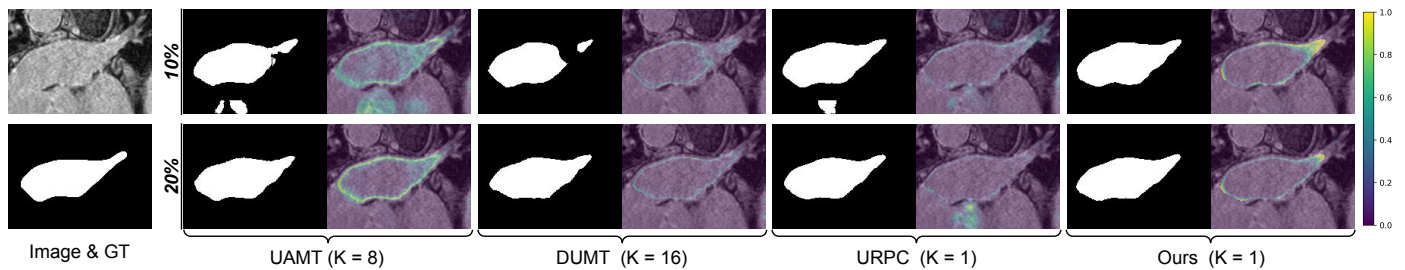
Figure 8: **Uncertainty analysis on the left atrium dataset** - Prediction and uncertainty map (overlaid on its image) are shown for each uncertainty-based method. The number of inferences for generating the uncertainty map is denoted as K.

Acknowledgements

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., et al., 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion 76, 243–297.

Adiga Vasudeva, S., Dolz, J., Lombaert, H., 2022. Leveraging labeling representations in uncertainty-based semi-supervised segmentation, in: Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 265–275.

Ayache, N., Duncan, J., 2016. 20th anniversary of the medical image analysis journal (MedIA). Medical Image Analysis 33, 1–3.

Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A., Matthews, P.M., Rueckert, D., 2017. Semi-supervised learning for network-based cardiac MR image segmentation, in: Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 253–260.

Baumgartner, C.F., Tezcan, K.C., Chaitanya, K., Hötker, A.M., Muehlematter, U.J., Schawkat, K., Becker, A.S., Donati, O., Konukoglu, E., 2019. PHiSeg: Capturing uncertainty in medical image segmentation, in: Medical Image Computing and Computer Assisted Intervention, Springer. pp. 119–127.

Bortsova, G., Dubost, F., Hogeweg, L., Katramados, I., Bruijne, M.d., 2019. Semi-supervised medical image segmentation via learning consistency under transformations, in: Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 810–818.

Camarasa, R., Bos, D., Hendrikse, J., Nederkoorn, P.J., Kooi, E., van der Lugt, A., de Bruijne, M., 2021. A quantitative comparison of epistemic uncertainty maps applied to multi-class segmentation. The Journal of Machine Learning for Biomedical Imaging 13, 1–39.

Chaitanya, K., Karani, N., Baumgartner, C.F., Becker, A., Donati, O., Konukoglu, E., 2019. Semi-supervised and task-driven data augmentation, in: International Conference on Information Processing in Medical Imaging, Springer. pp. 29–41.

Cheplygina, V., de Bruijne, M., Pluim, J.P., 2019. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical Image Analysis 54, 280–296.

Cui, W., Liu, Y., Li, Y., Guo, M., Li, Y., Li, X., Wang, T., Zeng, X., Ye, C., 2019. Semi-supervised brain lesion segmentation with an adapted mean teacher model, in: Information Processing in Medical Imaging, Springer. pp. 554–565.

Dalca, A.V., Guttag, J., Sabuncu, M.R., 2018. Anatomical priors in convolutional networks for unsupervised biomedical segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9290–9299.

Duncan, J.S., Ayache, N., 2000. Medical image analysis: Progress over two decades and the challenges ahead. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 85–106.

Durieux, P., Gevenois, P.A., Muylem, A.V., Howarth, N., Keyzer, C., 2018. Abdominal attenuation values on virtual and true unenhanced images obtained with third-generation dual-source dual-energy ct. American Journal of Roentgenology 210, 1042–1058.

Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: International Conference on Machine Learning, PMLR. pp. 1050–1059.

Grandvalet, Y., Bengio, Y., 2004. Semi-supervised learning by entropy minimization. Advances in Neural Information Processing Systems 17.

Huang, H., Chen, Q., Lin, L., Cai, M., Zhang, Q.W., Iwamoto, Y., Han, X., Furukawa, A., Kanasaki, S., Chen, Y.W., et al., 2022. MTL-ABS3Net: Atlas-based semi-supervised organ segmentation network with multi-task learning for medical images. IEEE Journal of Biomedical and Health Informatics .

Jiao, R., Zhang, Y., Ding, L., Cai, R., Zhang, J., 2022. Learning with limited annotations: A survey on deep semi-supervised learning for medical image segmentation. arXiv preprint arXiv:2207.14191 .

Judge, T., Bernard, O., Porumb, M., Chartsias, A., Beqiri, A., Jodoin, P.M., 2022. CRISP-reliable uncertainty estimation for medical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 492–502.

Karani, N., Erdil, E., Chaitanya, K., Konukoglu, E., 2021. Test-time adaptable neural networks for robust medical image segmentation. Medical Image Analysis 68, 101907.

Kendall, A., Badrinarayanan, V., Cipolla, R., 2017. Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. British Machine Vision Conference .

Kendall, A., Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision? Advances in Neural Information Processing Systems 30.

Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K., Eslami, S., Jimenez Rezende, D., Ronneberger, O., 2018. A probabilistic U-net for segmentation of ambiguous images. Advances in Neural Information Processing Systems 31.

Laine, S., Aila, T., 2017. Temporal ensembling for semi-supervised learning. International Conference on Learning Representations .

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems 30.

Larrazabal, A.J., Martínez, C., Glocker, B., Ferrante, E., 2020. Post-DAE: anatomically plausible segmentation via post-processing with denoising autoencoders. IEEE Transactions on Medical Imaging 39, 3813–3820.

Li, S., Zhang, C., He, X., 2020a. Shape-aware semi-supervised 3D semantic segmentation for medical images, in: Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 552–561.

Li, X., Yu, L., Chen, H., Fu, C.W., Xing, L., Heng, P.A., 2020b. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. IEEE Transactions on Neural Networks and Learning Systems 32, 523–534.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical Image Analysis 42, 60–88.

Loshchilov, I., Hutter, F., 2017. SGDR: stochastic gradient descent with warm restarts. International Conference on Learning Representations .

Luo, X., Chen, J., Song, T., Wang, G., 2021. Semi-supervised medical image segmentation through dual-task consistency, in: AAAI Conference on Artificial Intelligence, pp. 8801–8809.

Luo, X., Wang, G., Liao, W., Chen, J., Song, T., Chen, Y., Zhang, S., Metaxas, D.N., Zhang, S., 2022. Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. Medical Image Analysis 80, 102517.

Ma, J., Zhang, Y., Gu, S., An, X., Wang, Z., Ge, C., Wang, C., Zhang, F., Wang, Y., Xu, Y., et al., 2022. Fast and low-gpu-memory abdomen ct organ segmentation: The flare challenge. Medical Image Analysis 82, 102616.

Mehta, R., Filos, A., Baid, U., Sako, C., McKinley, R., Rebsamen, M., Dätwyler, K., Meier, R., Radojewski, P., Murugesan, G.K., et al., 2022. QU-BraTS: MICCAI BraTS 2020 challenge on quantifying uncertainty in brain tumor segmentation-analysis of ranking scores and benchmarking results. Journal of Machine Learning for Biomedical Imaging 1.

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

Monteiro, M., Le Folgoc, L., Coelho de Castro, D., Pawlowski, N., Marques, B., Kamnitsas, K., van der Wilk, M., Glocker, B., 2020. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. Advances in Neural Information Processing Systems 33, 12756–12767.

Neal, R.M., 2012. Bayesian learning for neural networks. volume 118. Springer Science & Business Media.

Nie, D., Gao, Y., Wang, L., Shen, D., 2018. ASDNet: attention based semi-supervised deep networks for medical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 370–378.

Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M., Bai, W., Caballero, J., Cook, S.A., De Marvao, A., Dawes, T., O'Regan, D.P., et al., 2017. Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation. IEEE Transactions on Medical Imaging 37, 384–395.

Painchaud, N., Skandarani, Y., Judge, T., Bernard, O., Lalande, A., Jodoin, P.M., 2020. Cardiac segmentation with strong anatomical guarantees. IEEE Transactions on Medical Imaging 39, 3703–3713.

Peng, J., Estrada, G., Pedersoli, M., Desrosiers, C., 2020. Deep co-training for semi-supervised image segmentation. Pattern Recognition 107, 107269.

Ravishankar, H., Venkataramani, R., Thiruvenkadam, S., Sudhakar, P., Vaidya, V., 2017. Learning and incorporating shape models for semantic segmentation, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 203–211.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training gans. Advances in Neural Information Processing Systems 29.

Sedai, S., Antony, B., Rai, R., Jones, K., Ishikawa, H., Schuman, J., Gadi, W., Garnavi, R., 2019. Uncertainty guided semi-supervised segmentation of retinal layers in OCT images, in: Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 282–290.

Tarvainen, A., et al., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in Neural Information Processing Systems 30.

Van Waerebeke, M., Lodygensky, G., Dolz, J., 2022. On the pitfalls of entropy-based uncertainty for multi-class semi-supervised segmentation. Uncertainty for Safe Utilization of Machine Learning in Medical Imaging Workshop in Medical Image Computing and Computer-Assisted Intervention .

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A., Bottou, L., 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research 11.

Van der Walt, S., Schönberger, J.L., Nunez-Iglesias, J., Boulogne, F., Warner, J.D., Yager, N., Gouillart, E., Yu, T., 2014. Scikit-image: image processing in python. PeerJ 2, e453.

Wang, K., Zhan, B., Zu, C., Wu, X., Zhou, J., Zhou, L., Wang, Y., 2021. Tripled-uncertainty guided mean teacher model for semi-supervised medical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 450–460.

Wang, K., Zhan, B., Zu, C., Wu, X., Zhou, J., Zhou, L., Wang, Y., 2022. Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning. Medical Image Analysis 79, 102447.

Wang, Y., Zhang, Y., Tian, J., Zhong, C., Shi, Z., Zhang, Y., He, Z., 2020. Double-uncertainty weighted method for semi-supervised learning, in: Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 542–551.

Wu, J., Fan, H., Zhang, X., Lin, S., Li, Z., 2021. Semi-supervised semantic segmentation via entropy minimization, in: International Conference on Multimedia and Expo, IEEE. pp. 1–6.

Wu, Y., Ge, Z., Zhang, D., Xu, M., Zhang, L., Xia, Y., Cai, J., 2022. Mutual consistency learning for semi-supervised medical image segmentation. Medical Image Analysis 81, 102530.

Xia, Y., Liu, F., Yang, D., Cai, J., Yu, L., Zhu, Z., Xu, D., Yuille, A., Roth, H., 2020. 3D semi-supervised learning with uncertainty-aware multi-view co-training, in: IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3646–3655.

Xiong, Z., Xia, Q., Hu, Z., Huang, N., Bian, C., Zheng, Y., Vesal, S., Ravikumar, N., Maier, A., Yang, X., et al., 2021. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. Medical Image Analysis 67, 101832.

Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A., 2019. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation, in: Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 605–613.

Zheng, H., Lin, L., Hu, H., Zhang, Q., Chen, Q., Iwamoto, Y., Han, X., Chen, Y.W., Tong, R., Wu, J., 2019. Semi-supervised segmentation of liver using adversarial learning with deep atlas prior, in: Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 148–156.

Zheng, H., Motch Perrine, S.M., Pitirri, M.K., Kawasaki, K., Wang, C., Richtsmeier, J.T., Chen, D.Z., 2020. Cartilage segmentation in high-resolution 3D micro-CT images via uncertainty-guided self-training with very sparse annotation, in: Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 802–812.