META-LEARNING FOR BOOTSTRAPPING MEDICAL IMAGE SEGMENTATION FROM IMPERFECT SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical imaging has witnessed remarkable progress but usually requires a large amount of high-quality annotated data which is time-consuming and costly to obtain. To alleviate the annotation burden, learning from imperfect supervision (scarce or noisy annotations) has received much attention recently. In this paper, we present **M**eta-**L**earning for **B**ootstrapping Medical Image **Seg**mentation (**MLB-Seg**), a unified meta-learning framework to sufficiently exploit the potential of imperfect supervision for medical image segmentation. In the face of noisy labeled data and unlabeled data, we first learn a segmentation model from a small clean set to generate initial labels for the unlabeled data and then gradually leverage the learner's own predictions (*i.e.*, the online pseudo labels) to bootstrap itself up via meta-learning. Specifically, MLB-Seg learns to dynamically assign per-pixel weight maps to both the imperfect labels (including both the generated labels and the noisy labels), as well as the pseudo labels commensurately to facilitate the bootstrapping procedure, where the weights are determined in a meta-process. To further improve the quality of the pseudo labels, we apply a consistency-based Pseudo Label Enhancement (PLE) scheme by ensembling predictions from various augmented versions of the same input. Noticing that the weight maps from these augmented variants can be extremely noisy from the meta-update, mean teacher is introduced into PLE to stabilize the weight map generation from the student (target) meta-learning model. Extensive experimental results on the public atrial and prostate segmentation datasets demonstrate that our method 1) achieves the state-of-the-art result under both semi- and noisy- supervision; 2) is robust against various imperfect supervisions. Code is publicly available at `https://anonymous.4open.science/r/MLB-Seg-C80E`.

1 INTRODUCTION

Reliable and robust segmentation of medical images plays a significant role in clinical diagnosis (Masood et al., 2015). In recent years, with the rapid development of deep learning, great success has been achieved in various image segmentation tasks (He et al., 2017; Long et al., 2015; Ronneberger et al., 2015; Zhou et al., 2018). However, training these models (Wang et al., 2017; Yuan et al., 2017; Zhao et al., 2018) requires large-scale image data with precise pixel-wise annotations, which are usually time-consuming and costly to obtain. In light of these challenges, recent studies have shifted the research focus to medical image segmentation with imperfect datasets (Tajbakhsh et al., 2020), such as datasets with scarce annotations where only limited annotated data is available for training, and noisy annotations (*e.g.*, partial and dilated annotation (Qu et al., 2020; Peng et al., 2019; Shi et al., 2021)) which is usually more easily accessible. Therefore, a more practical solution is to train a segmentation network under such imperfect supervision, which are more easily obtained while only using a small number of high-quality labeled samples (Wang et al., 2020a).

Specifically, in this paper, we only consider three types of imperfect supervision, *i.e.*, semi-, noisy-supervision, and a mixture of both in the context of medical image segmentation. Previous studies in this direction either only study semi-supervised settings (Peng et al., 2021; Yu et al., 2019a; Luo et al., 2020; Wu et al., 2021; Wang et al., 2020a), or only focus on noisy-supervised settings (Mirikharaji et al., 2019; Xue et al., 2019; Mirikharaji et al., 2019). To the best of our knowledge, in this paper, we develop the first general medical image segmentation framework to tackle both semi- and noisy-supervision at once. A naive solution is to first learn a segmentation model from a small clean

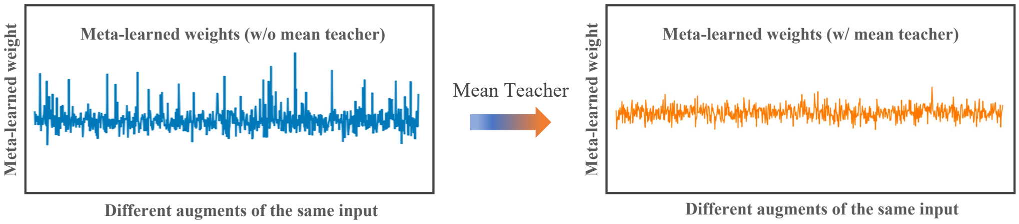


Figure 1: Illustration of the averaged meta-learned weights (by taking the mean value of each meta-learned weight map) of the augmented variants from the same input during one epoch w/ and w/o mean teacher.

set to generate initial labels for the unlabeled data and then gradually leverage the learner's own predictions (*i.e.*, the online pseudo labels) to bootstrap itself up (Reed et al., 2014). However, the error occurring in the generated initial labels and the noisy labels will be reinforced by the network during the bootstrapping procedure.

To alleviate this negative effect, our key idea is to use an adaptive reweighting scheme which learns to assign per-pixel weight to both the imperfect labels (including both generated initial labels and noisy labels), as well as the pseudo labels commensurately to facilitate the bootstrapping procedure. To achieve this goal, we present **M**eta-**L**earning for **B**ootstrapping Medical Image **Seg**mentation (**MLB-Seg**), which learns spatially adaptive weight maps associated with training images in a meta-process and then dynamically adjust the importance weight of each pixel in the imperfect labels and the pseudo labels for training. For each training iteration, higher importance will be assigned to pixels whose loss gradient direction is closer to those of clean data to approximate the optimal weight maps of the current batch based on the loss on a small set of clean images annotated by experts. For instance, the model will automatically assign higher importance to those pixels with correct pseudo labels when the imperfect labels are corrupted and vice versa.

To further improve the quality of the pseudo labels, we apply a consistency-based Pseudo Label Enhancement (PLE) scheme which ensembles predictions from various augmented versions of the same input by enforcing the consistency among them, similar to (Berthelot et al., 2019; Sohn et al., 2020). However, surprisingly, we observe that this strategy does not necessarily improve performance, but rather decreases the segmentation accuracy after the number of augmentations increases (see Section 4.5). This counter-intuitive observation motivates us to take a closer look at the weight updating procedure in the meta-process. As shown in Figure 1, we find that even for the same input, weight maps from different augmented variants can be extremely noisy from the meta-update using the same checkpoint. We hypothesize that this is due to that for some training samples with high pixel corruption ratios, introducing more augmented variants can also greatly enhance the noise as well, therefore leading to instability for the corresponding weight update. To address this issue, we introduce a mean-teacher model (Tarvainen & Valpola, 2017) into PLE for stabilizing the weight map generation from the student meta-learning model. Benefiting from the consistency-based weight ensembling mechanism, the mean-teacher model successfully stabilizes the meta-learned weight when largely increasing the number of augmentations.

The technical contributions of this paper are summarized as follows:

- To the best of our knowledge, we present the first unified meta-learning framework, MLB-Seg, for concurrently exploiting semi- and noisy- supervision in medical image segmentation. We also construct benchmarks based on LA (Chen et al., 2018) and PROMISE12 dataset (Litjens et al., 2014) by splitting all data into three subsets (*i.e.* clean, noisy, unlabeled) and synthesizing diverse types of noise for the noisy subset.
- MLB-Seg learns spatially adaptive weight maps in a meta-process to dynamically adjust the per-pixel importance in the imperfect and pseudo labels, for bootstrapping the segmentation model.
- To ensure the reliability of pseudo labels, MLB-Seg deploys a consistency-based Pseudo Label Enhancement (PLE) scheme. We identify an instability issue with the meta-learned weights by using PLE and propose to use a mean teacher model to stabilize the weights.
- Our method achieves the state-of-the-art result on the public LA and PROMISE12 benchmarks. Extensive experiments and component analysis demonstrate the effectiveness of our method under various imperfect supervision settings.

2 Related work

Semi-Supervised Learning. Recently, significant progress has been made in semi-supervised learning which could be used to deal with a large amount of unlabeled data. Pseudo label generation and consistency regularization are two popular methods used in SSL. (Xie et al., 2020; Ouali et al., 2020; Ke et al., 2019) are consistency-based methods. (Lee et al., 2013; Bachman et al., 2014) aim to generate pseudo labels for unlabeled data. Mixmatch (Berthelot et al., 2019) combines both consistency regularization and pseudo labelling. Fixmatch (Sohn et al., 2020) then utilize both weak and strong augmentation to get better improvements. FlexMatch (Zhang et al., 2021) further takes both learning status and different classes into consideration based on the curriculum learning approach. Particularly for the semi-supervised segmentation task, in (French et al., 2019), CutMix augmentation is adapted for semi-supervised semantic segmentation. CCT (Ouali et al., 2020) proposes a cross-consistency training algorithm to force the outputs to be consistent after introducing perturbations into the network. In (Peng et al., 2021), a self-paced strategy is applied to learn from noisy labels assisted by meta-label under curriculum learning. AEL (Hu et al., 2021) adaptively balances the training of well and badly performed categories, with a confidence bank to dynamically track category-wise performance during training. Besides the abundant applications for unlabeled data, semi-supervised learning could also be applied for noisy labels. DivideMix (Li et al., 2020a) introduces a dynamic way to split the training set into the labeled and unlabeled sets and train them in a semi-supervised way. ELR (Liu et al., 2020) introduces a regularization term to attend to the early-learning phenomenon and prevent bad influence brought by false labels. PES (Bai et al., 2021) proposes a progressive early stopping method to train different layers with different early stopping epochs.

Noisy-Supervised Learning. To alleviate the label noise issue, one popular line of methods learns to reweight the training loss using meta-learning. For instance, Zhu *et al.* proposed an end-to-end network to evaluate the label quality and reweight the corresponding sample-wise loss according to the evaluation scores (Zhu et al., 2019). Xue *et al.* (Xue et al., 2019) leveraged meta-learning to reweight different samples for more robust skin lesion classification. Zhang *et al.* (Zhang & Sabuncu, 2018) combines cross-entropy loss with Mean Absolute Error while using softmax outputs as the sample weights. Other research works extend these sample-wise reweighting methods to further perform pixel-wise reweighting. For instance, Mirikharaji *et al.* (Mirikharaji et al., 2019) generated the per-pixel importance weight based on the pixel-wise loss gradient direction to adjust the contribution of each pixel to model optimization but weight maps are only learned for noisy labels which fails to take the advantage of the complementarity between noisy labels and pseudo labels. Similarly, Wang *et al.* proposed the MCPM method (Wang et al., 2020a) which generates the weight map by feeding the loss value map into the meta-model. However, the loss reweighting-based methods attend less to a subset of the training data, which can result in information loss, especially regarding those noisy pixels. Other methods like Co-teaching (Han et al., 2018) trains two networks simultaneously aiming to teach each other the selected useful information obtained from its input. Yu *et al.* (Yu et al., 2019b) further propose Co-teaching+ that follows the idea of (Han et al., 2018) but only keeps the disagreement predicted data between two networks for parameters update. Ma *et al.* (Ma et al., 2020) introduce a simple normalization for constructing more robust loss.

3 Method

We present MLB-Seg, a unified meta-learning based medical image segmentation method which exploits a small clean set to learn more robust representation under imperfect supervision. Specifically, we consider semi-, noisy- supervision and a mixure of both throughout this paper. Assume that we are given a training dataset consisting of clean data $D_c = \{(x_i^c, y_i^c)\}_{i=1}^{N}$, noisy data $D_n = \{(x_j^n, y_j^n)\}_{j=1}^{M}$ and unlabeled data $D_u = \{(x_k^u)\}_{k=1}^{K}$ ($N \ll M + K$ and either M or K could be zero but not both), where the input image x_i^c, x_j^n, x_k^u are of size $H \times W$ with the corresponding clean ground-truth mask y_i^c and the corrupted noisy mask y_j^n.

3.1 Meta-learning for Bootstrapping

We first estimate labels for all unlabeled data using the baseline model which are trained on the clean data, denoted as follows

$$y_k^{\tilde{n}} = f_c(x_k^u; \theta_c), \tag{1}$$

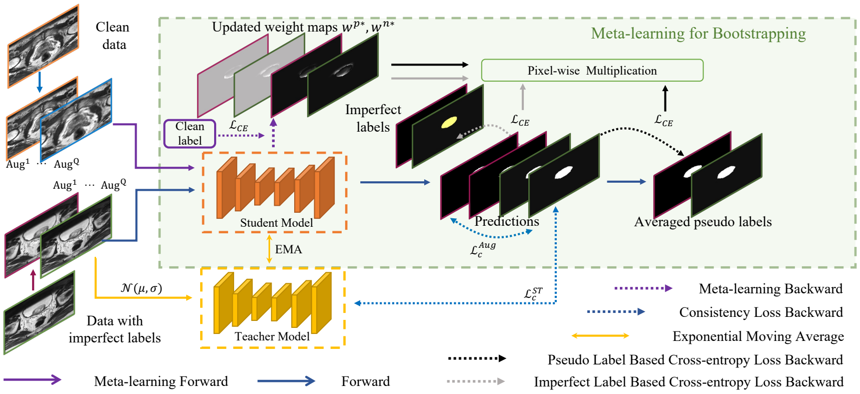

Clean data
Aug^1 ··· Aug^Q
Aug^1 ··· Aug^Q
Data with imperfect labels
$\mathcal{N}(\mu,\sigma)$
Meta-learning for Bootstrapping
Updated weight maps w^{p*}, w^{n*}
Clean label
$\mathcal{L}_{CE}$
Imperfect labels
Pixel-wise Multiplication
$\mathcal{L}_{CE}$
$\mathcal{L}_{CE}$
Student Model
EMA
Teacher Model
Predictions
$\mathcal{L}_c^{Aug}$
$\mathcal{L}_c^{ST}$
Averaged pseudo labels
Meta-learning Backward
Consistency Loss Backward
Exponential Moving Average
Pseudo Label Based Cross-entropy Loss Backward
Imperfect Label Based Cross-entropy Loss Backward
Meta-learning Forward
Forward

Figure 2: Schematic of the proposed MLB-Seg. Weight maps w^{n*}, w^{p*} associated with the imperfect labels and pseudo labels are meta-learned and optimization is further improved by enhancing the pseudo label estimation. A mean teacher model is used provide guidance for stabilizing the weight meta-update in the student model.

where $f_c(:;\theta_c)$ denotes the trained model parameterized by θ_c and $k = 1, 2, ..., K$. Together with the existing noisy labeled data, we denote them as $\widetilde{D_n} = \{(x_j^{\tilde{n}}, y_j^{\tilde{n}})\}_{j=1}^{M+K}$.

We then develop a novel meta-learning model for medical image segmentation, which learns from the clean set D_c to bootstrap itself up by leveraging the learner's own predictions (*i.e.*, pseudo labels), called Meta-Learning for Bootstrapping (MLB). As shown in Fig. 2, by adaptively adjusting the contribution between the imperfect and pseudo labels commensurately in the loss function, our method effectively alleviates the negative effects from the erroneous pixels. Specifically, at training step t, given a training batch from $\widetilde{D_n}$ that $S^n = \{(x_j^{\tilde{n}}, y_j^{\tilde{n}}), 1 \leq j \leq b_n\}$ and a clean training batch $S^c = \{(x_i^c, y_i^c), 1 \leq i \leq b_c\}$ where b_n, b_c are the batch size respectively. Our objective is:

$$\theta^*(w^n, w^p) = \arg\min_{\theta} \sum_{j=1}^{M+K} w_j^n \circ \mathcal{L}(f(x_j^{\tilde{n}};\theta), y_j^{\tilde{n}}) + w_j^p \circ \mathcal{L}(f(x_j^{\tilde{n}};\theta), y_j^p), \tag{2}$$

$$y_j^p = \arg\max_{c=1,\ldots,C} f(x_j^{\tilde{n}};\theta), \tag{3}$$

where y_j^p is the pseudo label generated by $f(x_j^{\tilde{n}};\theta)$, $\mathcal{L}(\cdot)$ is the cross-entropy loss function, C is the number of classes (we set $C = 2$ throughout this paper), $w_j^n, w_j^p \in \mathbb{R}^{H \times W}$ are the weight maps used for adjusting the contribution between the imperfect and the pseudo labels in two different loss terms. $\circ$ denotes the Hadamard product. We aim to solve for Eq. 2 following 3 steps:

- **Step 1:** Update $\hat{\theta}_{t+1}$ based on S^n and current weight map set. Following (Ren et al., 2018), during training step t, we calculate $\hat{\theta}_{t+1}$ to approach the optimal $\theta^*(w^n, w^p)$ as follows:

$$\hat{\theta}_{t+1} = \theta_t - \alpha\nabla(\sum_{j=1}^{b_n} w_j^n \circ \mathcal{L}(f(x_j^{\tilde{n}};\theta), y_j^{\tilde{n}}) + w_j^p \circ \mathcal{L}(f(x_j^{\tilde{n}};\theta), y_j^p))\big|_{\theta=\theta_t}, \tag{4}$$

  where α represents the step size.

- **Step 2:** Generate the meta-learned weight maps w^{n*}, w^{p*} based on S^c and $\hat{\theta}_{t+1}$ by minimizing the standard cross-entropy loss in the meta-objective function over the clean training data:

$$w^{n*}, w^{p*} = \arg\min_{w^n, w^p \geq \mathbf{0}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f(x_i^c;\theta^*(w^n, w^p)), y_i^c). \tag{5}$$

  Note that here we restrict every element in $w^{n/p}$ to be non-negative to prevent potentially unstable training (Ren et al., 2018). Such meta-learned process yields weight maps which can better balance the contribution of the imperfect and the pseudo labels, thus reducing the negative effects brought by the erroneous pixels. Following (Ren et al., 2018; Zhou et al., 2022), we only apply one step gradient descent of $w_j^{n/p}$ based on a small clean-label data set S^c, to reduce the computational expense. To be specific, at training step t, $w^{n/p}_j$ is first initialized as $\mathbf{0}$, then we estimate w^{n*}, w^{p*} as:

$$(w^n_{j,t}, w^p_{j,t}) = -\beta\nabla(\frac{1}{b_c}\sum_{i=1}^{b_c}\mathcal{L}(f(x^c_i;\hat{\theta}_{t+1}), y^c_i))\big|_{w^n_j, w^p_j=\mathbf{0}}, \tag{6}$$

$$\widetilde{w}^{n_{r,s}}_{j,t} = \max(w^{n_{r,s}}_{j,t}, 0), \quad \widetilde{w}^{p_{r,s}}_{j,t} = \max(w^{p_{r,s}}_{j,t}, 0), \tag{7}$$

$$\widetilde{w}^{n_{r,s}}_{j,t} = \frac{\widetilde{w}^{n_{r,s}}_{j,t}}{\sum_{j=1}^{b_n}\sum_{r,s}\widetilde{w}^{n_{r,s}}_{j,t}+\epsilon}, \quad \widetilde{w}^{p_{r,s}}_{j,t} = \frac{\widetilde{w}^{p_{r,s}}_{j,t}}{\sum_{j=1}^{b_n}\sum_{r,s}\widetilde{w}^{p_{r,s}}_{j,t}+\epsilon}, \tag{8}$$

where β stands for the step size and $w^{n/p_{r,s}}_{j,t}$ indicates the value at r^{th} row, s^{th} column of $w^{n/p}_j$ at time t. Eq. 7 is used to enforce all weights to be strictly non-negative. Then Eq. 8 is introduced to normalize the weights in a single training batch so that they sum up to one. Here, we add a small number ϵ to keep the denominator greater than 0.

- **Step 3:** The meta-learned weight maps are used to spatially modulate the pixel-wise loss to update θ_{t+1}:

$$\theta_{t+1} = \theta_t - \alpha\nabla(\sum_{j=1}^{b_n}\widetilde{w}^n_{j,t}\circ\mathcal{L}(f(x^{\tilde{n}}_j;\theta), y^{\tilde{n}}_j) + \widetilde{w}^p_{j,t}\circ\mathcal{L}(f(x^{\tilde{n}}_j;\theta), y^p_j))\big|_{\theta=\theta_t}. \tag{9}$$

3.2 Consistency-based Pseudo Label Enhancement

To generate more reliable pseudo labels, we propose a consistency based Pseudo Label Enhancement (PLE) scheme by enforcing predictions across augmented versions of the same input to be consistent. Specifically, given Q augmentations of the same input image, we enhance the pseudo label by averaging the outputs of Q augmentations and the original input:

$$\widehat{y}^p_j = \operatorname*{arg\,max}_{c=1,\dots,C}\frac{1}{Q+1}(\sum_{q=1}^{Q}\tau_q^{-1}(f(x^{\tilde{n}^q}_j;\theta)) + f(x^{\tilde{n}^0}_j;\theta)), \tag{10}$$

where $f(x^{\tilde{n}^q}_j;\theta)$ is the output of q-th augmented sample, $f(x^{\tilde{n}^0}_{N+j};\theta)$ is the output of the original input, and τ_q^{-1} means the corresponding inverse transformation of the q-th augmented sample. Meanwhile, to further increase the output consistency among all the augmented samples and original input, we use an additional consistency loss $\mathcal{L}^{Aug}_c$ in the learning objective:

$$\mathcal{L}^{Aug}_c(x^{\tilde{n}}_j) = \frac{2}{(Q+1)Q}\frac{1}{HW}\sum_{q,v}\sum_{r,s}||f(x^{\tilde{n}^q}_j;\theta)_{r,s} - \tau_q(\tau_v^{-1}(f(x^{\tilde{n}^v}_j;\theta)))_{r,s}||^2, \tag{11}$$

where (r, s) denotes the pixel index. τ_q is the corresponding transformation to generate the q-th augmented sample. (q, v) denotes the pairwise combination among all augmented samples and the original input. The final loss is average mean square distance among all $\frac{(Q+1)Q}{2}$ pairs of combinations.

3.3 Mean Teacher for Stabilizing Meta-Learned Weights

Surprisingly, we observe that simply using PLE can result in great performance degradation after the number of augmentations increases (see Section 4.5). This counter-intuitive observation motivates us to take a closer look at the weight updating procedure in the meta-process. As shown in Figure 1, we find that even for the same input, weight maps from different augmented variants can be extremely noisy from the meta-update using the same checkpoint. We hypothesize that this is due to that for some training samples with high pixel corruption ratios, applying various augmentations also augments the noise distribution in the original sample, therefore, leading to the unstable meta-update. Consequently, the network parameter update can become unstable as well, which will compound the instability issue during the following training iterations.

To address this issue, we adopt the idea of mean teacher (Tarvainen & Valpola, 2017) together with the consistency loss (Eq. 13) to smooth and stabilize the network weight update process and thus, significantly increase the robustness during the meta-learned process and help the network produce more reliable meta-learned weight maps. As shown in Fig. 1, we use a teacher network with the same architecture which is used to provide guidance for the student meta-learning model. We find that combining PLE with mean teacher largely diminishes the fluctuations of meta-learned weight maps and the resulted weight maps become more robust. As shown in Fig. 2, the student model is used in meta-learning while the teacher model is used for model weight ensemble where Exponential Moving Average (EMA) (Tarvainen & Valpola, 2017) is applied to update the teacher model based on a weighted average of the student model parameters and its own parameters. The consistency loss is then adopted to enforce the uniformity between the student and teacher model. To be specific, we first apply disturbance to the input of the student model which then becomes the input of the teacher model. Then a consistency loss $\mathcal{L}_c^{ST}$ is used to maximize the similarity between the outputs from the teacher model and student model. Such mechanism could also introduce more reliability to the student model while stabilizing the teacher model. Specifically, for each input $x_j^{\tilde{n}}$ in the batch S^n, then corresponding input of teacher model is

$$x_j^T = x_j^{\tilde{n}} + \gamma\mathcal{N}(\mu, \sigma), \tag{12}$$

where $\mathcal{N}(\mu, \sigma) \in \mathbb{R}^{H \times W}$ denotes the Gaussian distribution with μ as mean and σ as standard deviation. And γ is used to control the noise level. Then, the consistency loss is implemented based on pixel-wise mean squared error (MSE) loss:

$$\mathcal{L}_c^{ST}(x_j^{\tilde{n}}) = \frac{1}{HW}\sum_{r,s} ||f(x_j^{\tilde{n}}; \theta_t^S)_{r,s} - f(x_j^T; \theta_t^T)_{r,s}||^2. \tag{13}$$

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our proposed method on two different datasets under 3 different imperfect supervision settings: 1) semi-supervision, 2) noisy-supervision, and 3) a mixture of semi- and noisy- supervision, referred to as mix-supervision. Four metrics, including the Dice coefficient, Jaccard Index (JI), Hausdorff Distance (HD), Average Surface Distance (ASD) are employed for quantifying the segmentation performance. For semi-supervised segmentation, we report results on the left atrial (LA) dataset from 2018 Atrial Segmentation Challenge (Chen et al., 2018) as well as the Prostate MR Image Segmentation 2012 (PROMISE2012) dataset (Litjens et al., 2014). For noisy-supervised segmentation, we report results on the PROMISE2012 dataset (Litjens et al., 2014). For mix-supervision, we create benchmarks based on both LA and PROMISE12 dataset to evaluate our algorithm which will be described below.

- **LA dataset.** LA dataset includes 100 3D Gadolinium-Enhanced Magnetic Resonance Imaging with 3D Binary Masks of the Left Atrial Cavity. These scans have an isotropic resolution of $0.625 \times 0.625 \times 0.625mm^3$. Following the setting of (Yu et al., 2019a), in order to compare the segmentation performance of different methods better, all the scans were cropped centering at the heart region and also normalized as zero mean and unit variance. During experiments, we split the whole dataset into 80 scans for training and 20 scans for evaluation which is also exactly the same as (Yu et al., 2019a). Among training set, we randomly pick 8 scans (10% of training set) as the perfect labeled data. All 2D input images were resized to 144×144.
- **PROMISE12.** PROMISE12 dataset contains 50 3D transversal T2-weighted MR images of the prostate with manual binary prostate gland segmentation and is obtained from multiple centers with different acquisition protocols. During all experiments,we use 40 cases as the training set and 10 cases as the evaluation set. 3 out of the 40 training cases are randomly picked and used as the meta set and all images are resized to 144×144. Notice that all the splits are randomly selected. Specially in Table 1, to ensure the fair comparison with Self-Paced (Peng et al., 2021), we split the 10 cases from evaluation set into 4 cases as the validation set, and 6 as the test set. And we also conducted experiments with images resized to 256×256 which is used in Self-Paced. And following Self-Paced (Soerensen et al., 2021; Wang et al., 2021), we use the 2D slices along the axial view for training and testing. We took 2D images along the short-axis of 3D slices. For normalization, each scan is normalized based on the 1% and 99% percentiles of the intensity histogram before extracting 2D images following (Peng et al., 2021).

Implementation Details. All of our experiments are based on 2D images. We adopt UNet++ as our baseline. Network parameters are optimized by SGD setting the learning rate at 0.005, momentum to be 0.9 and weight decay as 0.0005. Exponential moving average (EMA) decay rate is set as 0.99 following (Tarvainen & Valpola, 2017). For the label generation process, we first train with all clean labeled data for 30 epochs with batch size set as 16. We then use the latest model to generate labels for unlabeled data. Next, we train our MLB-Seg for 100 epochs.

Noisy-Supervision Synthesizing. We adopt a combination of random rotation, erosion or dilation to generate noisy labels, following (Shi et al., 2021). To explore the effectiveness of our method under different noise levels, we generate 3 corrupted noisy label sets under noise levels L_1, L_2, L_3, where the average Dice coefficients are $\text{Dice}_{L_1} = 0.4148$, $\text{Dice}_{L_2} = 0.6206$, and $\text{Dice}_{L_3} = 0.8031$ (*i.e.*, the corrupted ratios are around 60% (L_1), 40% (L_2), and 20% (L_3)). Unless otherwise specified, all results are reported with L_2 noise level. For the clean dataset, we keep it the same as the meta set.

Mix-Supervision Synthesizing. We create benchmarks for mix-supervision based on PROMISE12 as well as LA dataset. To be specific, for PROMISE12 and LA, we keep using the same 3 (PROMISE12) and 8 (LA) training cases as the clean dataset as mentioned in dataset description and also set them the same as the cases in the meta set. And the rest 37 (PROMISE12) and 72 (LA) training cases are then randomly split into two subsets (*i.e.* noisy labeled and unlabeled dataset) according to a manually defined mixing ratio. All noisy labels in the noisy labeled dataset are the same as the corresponding corrupted labels in L_2 noise level that the corrupted ratio is around 40%.

4.2 Results under Semi-Supervision

To illustrate the effectiveness of MLB-Seg under different types of imperfect supervision, we first evaluate our method under semi-supervision. We compare our method with the baseline (UNet++(Zhou et al., 2018)) and previous semi-supervised methods on PROMISE12 (Table 1) and LA dataset (Table 2), including adversarial learning based methods (Zheng et al., 2019; Zhang et al., 2017b), mean teacher and consistency based methods (Perone & Cohen-Adad, 2018; Yu et al., 2019a; Hang et al., 2020; Wang et al., 2020b; Luo et al., 2020; Wu et al., 2021), a pre-training based strategy (Vu et al., 2019), a data augmentation based strategy (Zhang et al., 2017a), a co-training strategy (Peng et al., 2020), and contrastive learning based methods (Chaitanya et al., 2020; Peng et al., 2021). Note that for a fair comparison in Table 1, we also conduct experiments using Unet (Ronneberger et al., 2015) as the backbone which is used in Self-Paced (Peng et al., 2021) and we select the best checkpoint based on the validation set and report the results on the test set. As shown in Table 1, 2, our MLB-Seg outperforms recent state-of-the-art methods on both PROMISE12 (under different combinations of backbones and image sizes) and the LA dataset under almost all evaluation measures.

Table 1: Comparison with existing methods under semi-supervision on PROMISE12.

Method	Dice (%)↑
UNet++ (Zhou et al., 2018)	68.85
Entropy Min. (Vu et al., 2019)	53.47
Mix-up (Zhang et al., 2017a)	41.38
Adv. Traning (Zhang et al., 2017b)	61.58
Mean Teacher (Perone & Cohen-Adad, 2018)	52.96
Discrete MI (Peng et al., 2020)	47.77
Contrastive (Chaitanya et al., 2020)	61.15
Self-Paced (Peng et al., 2021) (Unet, 256)	74.47
MLB-Seg (Unet, 144)	76.41
MLB-Seg (Unet, 256)	76.15
MLB-Seg (Unet++, 144)	77.22
MLB-Seg (Unet++, 256)	**78.27**

4.3 Results under Noisy-Supervision

We further evaluate our method under noisy-supervision. We compare our method with the baseline (UNet++), UNet++ meta and previous methods in Table3. To ensure a fair comparison, we use UNet++ as the backbone architecture for all comparison methods. Specifically, UNet++ meta represents training exclusively on the clean data. We also compare with recent methods, including a meta-learning based method which only reweights the noisy training labels (Mirikharaji et al., 2019), a regularization based method (Zhang et al., 2017a) and semi-supervised learning based methods (Li et al., 2020a; Liu et al., 2020). For (Zhang et al., 2017a; Li et al., 2020a; Liu et al., 2020), we follow their public source codes and modify the loss term for segmentation tasks.

As a reference, we also train UNet++ on the entire training set using clean labels **(No Noise)**. As shown in 3, our method outperforms existing methods by a large margin under almost all evaluation measures except for ASD where we are the second-best. For instance, compared with (Mirikharaji et al., 2019), our method yields 4.19% performance improvement in Dice under the same noise

Table 2: Comparison with existing methods under semi-supervision on LA dataset.

Method	Dice (%)↑	JI (%)↑	HD (voxel)↓	ASD (voxel)↓
UNet++ Zhou et al. (2018)	81.33	70.87	14.79	4.62
DAP Zheng et al. (2019)	81.89	71.23	15.81	3.80
UA-MT Yu et al. (2019a)	84.25	73.48	13.84	3.36
SASSNet Li et al. (2020b)	87.32	77.72	9.62	2.55
LG-ER-MT Hang et al. (2020)	85.54	75.12	13.29	3.77
DUWM Wang et al. (2020b)	85.91	75.75	12.67	3.31
DTC Luo et al. (2020)	86.57	76.55	14.47	3.74
MC-Net Wu et al. (2021)	87.71	78.31	9.36	**2.18**
MLB-Seg	**88.69**	**79.86**	**8.99**	2.61

Table 3: Comparison with existing methods under noisy-supervision on PROMISE12.

Method	Dice (%)↑	JI (%)↑	HD (voxel)↓	ASD (voxel)↓
UNet++ (Zhou et al., 2018)	73.74	58.90	11.63	3.70
UNet++ meta	73.04	58.51	17.06	5.50
NL reweighting (Mirikharaji et al., 2019)	76.64	62.62	8.33	2.75
Mix-up (Zhang et al., 2017a)	69.18	53.78	13.25	4.56
DivideMix (Li et al., 2020a)	74.77	59.99	8.09	2.40
ELR* (Liu et al., 2020)	75.43	61.91	6.88	**2.02**
MLB-Seg - L_2	**80.83**	**68.10**	**6.68**	2.10
No Noise	83.72	72.23	5.54	2.04

Table 4: Ablation study on mixing ratio for mix-supervision based on PROMISE12 and LA dataset.

GT	Noisy	Unlabeled	PROMISE12		LA	
			Dice (%)↑	ASD (voxel)↓	Dice (%) ↑	ASD (voxel)↓
0%	0%	100%	76.68	2.64	88.69	2.61
0%	25%	75%	77.70	2.79	88.79	2.87
25%	0%	75%	79.02	2.37	88.36	2.61
0%	50%	50%	79.34	2.42	88.60	2.32
0%	75%	25%	80.15	2.52	89.28	2.57
25%	75%	0%	80.58	2.25	89.51	2.51
0%	100%	0%	80.83	2.10	89.17	2.35

level. Furthermore, we also investigate the robustness of our method by varying the noise level of the corrupted training set from $\{L_1, L_2, L_3\}$. At each noise level, we compare the baseline UNet++ which is directly trained on the noisy training data with our MLB-Seg (see Supplementary for the detailed Table). Compared with the baseline method which drops from 80.03% to 59.77%, when increasing the noise level from L_3 to L_1, our proposed method only drops from 82.01% to 77.70%. This promising result shows the robustness of our method and also reveals that under a more severe noise level, MLB-Seg delivers larger performance improvement.

4.4 Results under Mix-Supervision

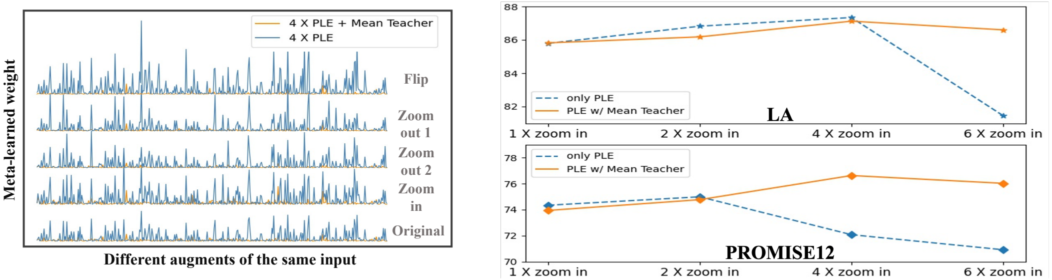
We also evaluate MLB-Seg under mix-supervision (including both semi- and noisy- supervision). Moreover, to investigate how the ratio of semi- and noisy- supervision will influence the performance of MLB-Seg, we conduct experiments with mixing ratios (noisy:unlabeled) at 25%:75%, 50%:50% and 75%:25% on both PROMISE12 and LA dataset. For ratios 25%:75% and 75%:25%, we also replace labels of the 25% part with ground-truth to test the upper bound performance. As shown in Table 4 (see Supplementary for more detailed Table), we can see that MLB-Seg is quite robust under different mixing ratios. For mixing ratios 25%:75% and 75%:25%, our MLB-Seg already yields a very high result with only a small gap compared with the corresponding upper bound (80.15% vs. 80.58% on PROMISE12 and 89.28% vs. 89.51% on LA).

4.5 Ablation Study

To explore how different components of our MLB-Seg contribute to the final result, we conduct the following experiments under semi-supervision on PROMISE12: 1) the bootstrapping method (Reed et al., 2014) (using fixed weights without applying meta-learning); 2) **MLB**, which only reweights the imperfect labels and pseudo labels without applying PLE and mean teacher; 3) **MLB + mean teacher** which combines MLB with mean teacher scheme; 4) **MLB + PLE** which applies PLE strategy with MLB. When applying multiple data augmentations (*i.e.*, for $Q = 2, 4$), we find the best performing combinations are 2 × PLE (using one zoom in and one zoom out), 4 × PLE (using one zoom in and two zoom out and one flip for each input); 4) **MLB + PLE + mean teacher** which combines PLE, mean teacher with MLB simultaneously to help better understand how mean teacher will contribute to PLE. Our results are summarized in Table 5, which shows the effectiveness of our proposed components. The best results are achieved when all components are used.

Table 5: Ablation study on different components used in MLB-Seg based on PROMISE12.

MLB	mean teacher	2× PLE	4× PLE	Dice (%)↑	JI (%)↑	HD (voxel)↓	ASD (voxel)↓
				70.85	55.85	10.02	4.35
✓				73.97	59.71	8.49	3.49
✓	✓			73.76	59.37	8.04	3.00
✓		✓		75.01	60.92	**7.58**	2.70
✓			✓	73.10	58.45	9.42	3.57
✓	✓		✓	**76.68**	**63.14**	7.85	**2.64**



(a) Average meta-learned weights of augmented variants from one input w/ and w/o mean teacher

(b) Comparison among of augmentations used in PLE w/ and w/o mean teacher on LA and PROMISE12 dataset

Figure 3: (a) Blue line denotes the average meta-learned weights of different augmented samples from one sample while using 4× PLE and orange line represents using 4× PLE w/ mean teacher. Applying different number of augmentations (zoom in) in PLE w/ and w/o mean teacher.

To demonstrate how PLE combined with the mean teacher model help stabilize the meta-weight update, we plot the number of augmentations (Q) vs. the performance of MLB + PLE (w/ mean teacher) and MLB + PLE + mean teacher on both LA and PROMISE12 dataset. Specifically, we apply 1 × zoom in, 2 × zoom in, 4 × zoom in and 6 × zoom in respectively for each input sample. As shown in Fig. 3b, we can see for MLB + PLE (w/ mean teacher) (the blue line) indicates that while increasing Q from 1 to 2, the performance does improve from 85.79% to 86.82% (LA) and 74.34% to 74.99% (PROMISE12). However, when Q reaches 6 (for LA) or Q reaches 4 and 6 (for PROMISE12), the performance significant drops from 87.34% to 81.46% (LA, $Q = 6$), from 74.99% to 72.07% (PROMISE12, $Q = 4$) and from 74.99% to 70.91% (PROMISE12, $Q = 6$) . However, for MLB + PLE + mean teacher (both orange lines), the performance reaches 76.63% (from 72.07%, PROMISE12) when $Q = 4$, 75.84% (from 70.91%, PROMISE12) and 86.60% (from 81.46%, LA) when $Q = 6$, which suggests that deploying the mean teacher model combined with PLE indeed helps stabilizes the meta-learning model. When $Q \geq 4$, the performance of MLB + PLE + mean teacher begins to saturate while that of MLB + PLE (w/ mean teacher) keeps decreasing. In Fig. 3a, using different augmentation types, we observe that MLB + PLE + mean teacher consistently improves the stability of meta-learning compared with its counterpart without using mean teacher, further validating the effectiveness of our method. Based on our ablation results, we adopt **MLB + 4 × PLE + mean teacher** as our default configuration of MLB-Seg.

Qualitative Analysis. We also demonstrate a set of qualitative examples to illustrate how the proposed MLB-Seg benefits medical image segmentation in the Supplementary. We observe that MLB-Seg generally pays more attention to the edge information. For those erroneous pixels, higher weights are assigned to the pseudo labeled pixels while such pixels are assigned with nearly zero in the imperfect label, which effectively alleviates the negative effects from erroneous pixels.

5 Conclusion

In this paper, we present a novel meta-learning based segmentation method for medical image segmentation under imperfect supervision. With few expert-level labels as guidance, our model bootstraps itself up by dynamically reweighting the contributions from imperfect labels and its own outputs, thus alleviating the negative effects from the erroneous voxels. In addition, we identify an instability issue with the meta-learned weights when using data augmentation and propose to use a mean teacher model to stabilize the weights accordingly. Extensive experiments demonstrate that our method is not only effective under various imperfect supervision settings but also quite robust against various noise levels without sacrificing the segmentation accuracy. Notably, our method achieves the state-of-the-art result on both LA and PROMISE12 benchmarks. We hope our effort will encourage future exploration on leveraging the widely existing imperfect medical datasets.

REFERENCES

Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27, 2014.

Yingbin Bai, Erkun Yang, Bo Han, Yanhua Yang, Jiatong Li, Yinian Mao, Gang Niu, and Tongliang Liu. Understanding and improving early stopping for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34, 2021.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in Neural Information Processing Systems*, 33:12546–12558, 2020.

Chen Chen, Wenjia Bai, and Daniel Rueckert. Multi-task learning for left atrial segmentation on ge-mri. In *International workshop on statistical atlases and computational models of the heart*, pp. 292–301. Springer, 2018.

Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv:1906.01916*, 2019.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.

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

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui, and Liwei Wang. Semi-supervised semantic segmentation via adaptive equalization learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Zhanghan Ke, Daoye Wang, Qiong Yan, Jimmy Ren, and Rynson WH Lau. Dual student: Breaking the limits of the teacher in semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6728–6736, 2019.

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, pp. 896, 2013.

Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020a.

Shuailin Li, Chuyu Zhang, and Xuming He. Shape-aware semi-supervised 3d semantic segmentation for medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 552–561. Springer, 2020b.

Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342, 2020.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. *arXiv preprint arXiv:2009.04448*, 2020.

Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *International Conference on Machine Learning*, pp. 6543–6553. PMLR, 2020.

Saleha Masood, Muhammad Sharif, Afifa Masood, Mussarat Yasmin, and Mudassar Raza. A survey on medical image segmentation. *Current Medical Imaging*, 11(1):3–14, 2015.

Zahra Mirikharaji, Yiqi Yan, and Ghassan Hamarneh. Learning to segment skin lesions from noisy annotations. In *Domain adaptation and representation transfer and medical image learning with less labels and imperfect data*, pp. 207–215. Springer, 2019.

Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12674–12684, 2020.

Jizong Peng, Guillermo Estrada, Marco Pedersoli, and Christian Desrosiers. Deep co-training for semi-supervised image segmentation. *Pattern Recognition*, 107:107269, 2020.

Jizong Peng, Ping Wang, Christian Desrosiers, and Marco Pedersoli. Self-paced contrastive learning for semi-supervised medical image segmentation with meta-labels. *Advances in Neural Information Processing Systems*, 34, 2021.

Liying Peng, Lanfen Lin, Hongjie Hu, Yue Zhang, Huali Li, Yutaro Iwamoto, Xian-Hua Han, and Yen-Wei Chen. Semi-supervised learning for semantic segmentation of emphysema with partial annotations. *IEEE Journal of Biomedical and Health Informatics*, 24(8):2327–2336, 2019.

Christian S Perone and Julien Cohen-Adad. Deep semi-supervised segmentation with weight-averaged consistency targets. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pp. 12–19. Springer, 2018.

Hui Qu, Pengxiang Wu, Qiaoying Huang, Jingru Yi, Zhennan Yan, Kang Li, Gregory M Riedlinger, Subhajyoti De, Shaoting Zhang, and Dimitris N Metaxas. Weakly supervised deep nuclei segmentation using partial points annotation in histopathology images. *IEEE transactions on medical imaging*, 39(11):3655–3666, 2020.

Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, pp. 4334–4343. PMLR, 2018.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Jiangbo Shi, Chang Jia, Zeyu Gao, Tieliang Gong, Chunbao Wang, and Chen Li. Meta mask correction for nuclei segmentation in histopathological image. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2059–2063. IEEE, 2021.

Simon John Christoph Soerensen, Richard E Fan, Arun Seetharaman, Leo Chen, Wei Shao, Indrani Bhattacharya, Yong-hun Kim, Rewa Sood, Michael Borre, Benjamin I Chung, et al. Deep learning improves speed and accuracy of prostate gland segmentations on magnetic resonance imaging for targeted biopsy. *The Journal of Urology*, 206(3):604–612, 2021.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33:596–608, 2020.

Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63:101693, 2020.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2517–2526, 2019.

Guotai Wang, Wenqi Li, Sébastien Ourselin, and Tom Vercauteren. Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. In *International MICCAI brainlesion workshop*, pp. 178–190. Springer, 2017.

Jixin Wang, Sanping Zhou, Chaowei Fang, Le Wang, and Jinjun Wang. Meta corrupted pixels mining for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 335–345. Springer, 2020a.

Kaiping Wang, Bo Zhan, Chen Zu, Xi Wu, Jiliu Zhou, Luping Zhou, and Yan Wang. Tripled-uncertainty guided mean teacher model for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 450–460. Springer, 2021.

Yixin Wang, Yao Zhang, Jiang Tian, Cheng Zhong, Zhongchao Shi, Yang Zhang, and Zhiqiang He. Double-uncertainty weighted method for semi-supervised learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 542–551. Springer, 2020b.

Yicheng Wu, Minfeng Xu, Zongyuan Ge, Jianfei Cai, and Lei Zhang. Semi-supervised left atrium segmentation with mutual consistency training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 297–306. Springer, 2021.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268, 2020.

Cheng Xue, Qi Dou, Xueying Shi, Hao Chen, and Pheng-Ann Heng. Robust learning at noisy labeled medical images: Applied to skin lesion classification. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 1280–1283. IEEE, 2019.

Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 605–613. Springer, 2019a.

Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *International Conference on Machine Learning*, pp. 7164–7173. PMLR, 2019b.

Yading Yuan, Ming Chao, and Yeh-Chi Lo. Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance. *IEEE transactions on medical imaging*, 36(9): 1876–1886, 2017.

Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34, 2021.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017a.

Yizhe Zhang, Lin Yang, Jianxu Chen, Maridel Fredericksen, David P Hughes, and Danny Z Chen. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In *International conference on medical image computing and computer-assisted intervention*, pp. 408–416. Springer, 2017b.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

Xiaomei Zhao, Yihong Wu, Guidong Song, Zhenye Li, Yazhuo Zhang, and Yong Fan. A deep learning model integrating fcnns and crfs for brain tumor segmentation. *Medical image analysis*, 43:98–111, 2018.

Han Zheng, Lanfen Lin, Hongjie Hu, Qiaowei Zhang, Qingqing Chen, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, Ruofeng Tong, and Jian Wu. Semi-supervised segmentation of liver using adversarial learning with deep atlas prior. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 148–156. Springer, 2019.

Yuyin Zhou, Xianhang Li, Fengze Liu, Xuxi Chen, Lequan Yu, Cihang Xie, Matthew P Lungren, and Lei Xing. Learning to bootstrap for combating label noise. *arXiv preprint arXiv:2202.04291*, 2022.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pp. 3–11. Springer, 2018.

Haidong Zhu, Jialin Shi, and Ji Wu. Pick-and-learn: automatic quality evaluation for noisy-labeled image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 576–584. Springer, 2019.