DSAL: Deeply Supervised Active Learning from Strong and Weak Labelers for Biomedical Image Segmentation

Ziyuan Zhao, Zeng Zeng, *Senior Member, IEEE*, Kaixin Xu, Cen Chen, Cuntai Guan, *Fellow, IEEE*

Abstract—Image segmentation is one of the most essential biomedical image processing problems for different imaging modalities, including microscopy and X-ray in the Internet-of-Medical-Things (IoMT) domain. However, annotating biomedical images is knowledge-driven, time-consuming, and labor-intensive, making it difficult to obtain abundant labels with limited costs. Active learning strategies come into ease the burden of human annotation, which queries only a subset of training data for annotation. Despite receiving attention, most of active learning methods still require huge computational costs and utilize unlabeled data inefficiently. They also tend to ignore the intermediate knowledge within networks. In this work, we propose a deep active semi-supervised learning framework, DSAL, combining active learning and semi-supervised learning strategies. In DSAL, a new criterion based on deep supervision mechanism is proposed to select informative samples with high uncertainties and low uncertainties for strong labelers and weak labelers respectively. The internal criterion leverages the disagreement of intermediate features within the deep learning network for active sample selection, which subsequently reduces the computational costs. We use the proposed criteria to select samples for strong and weak labelers to produce oracle labels and pseudo labels simultaneously at each active learning iteration in an ensemble learning manner, which can be examined with IoMT Platform. Extensive experiments on multiple medical image datasets demonstrate the superiority of the proposed method over state-of-the-art active learning methods.

Index Terms—Biomedical Image Segmentation, Active Learning, Semi-supervised Learning, Ensemble Learning, Internet of Medical Things

I. INTRODUCTION

IMAGE segmentation plays an important role in the Internet of Medical things (IoMT) systems and different applications in biology and medicine [1]–[3]. The information from different sources including medical devices, models, doctors and sensors in the IoMT are linked by the internet or cloud

services to optimize the processes for consultation, diagnosis and follow-up [4]. In IoMT systems, image segmentation is a vital prerequisite of computer-aided diagnosis (CAD). The past few decades have witnessed great developments in segmentation models based on deep learning and Convolutional Neural Networks (CNNs) [5]–[8] for diverse imaging modalities, such as microscopy and X-ray. However, deep learning methods heavily rely on high volumes of annotated training samples, which are difficult to obtain, especially for biomedical imaging applications. Extensive domain knowledge from biomedical experts is required for manual annotation and inspection. Besides, annotating biomedical data is costly and laborious, which results in the fact that rarely we have abundant labeled samples for biomedical image segmentation.

In order to tackle the paucity of labeled data, some approaches different from traditional supervised learning have been developed [9]. Utilizing unlabeled data, semi-supervised learning (SSL) methods [10] involve a self-training process to produce pseudo labels, in which, model updates and pseudo annotations, given the labeled data and model parameters, are performed in an alternating manner. Using unlabeled data in SSL leads to further improvement in model performance. However, there is no self-correcting process in SSL, which may cause inaccurate predictions and tend to find suboptimal solutions in the training process. Yet, a small set of accurate labels is more effective than quantities of inaccurate ones. By doing this, active learning (AL) paradigms can be explored to select valuable samples to be labeled with high quality [11].

AL strategies allow enlarging the training dataset by iteratively selecting informative samples for expert annotation [12]. To minimize the involvement of experts, various effective selection criteria defined on model predictions are used to discover valuable samples, which are expected to maximally boost the model performance once their annotations are obtained. AL has been extensively studied on various vision tasks [13]. However, there are some barriers that impede its application in clinical practice. First, the repetitive and extensive training process can harm the interactive annotation process. Second, some methods involve optimization theory and methods with high computational complexity, which further increase time costs. Third, the labeling costs are vastly different across examples, but each example is assumed to be equally expensive for the experts in most work, which leads to inefficiency and redundancy in data annotation [14].

In this work, we aim to alleviate the aforementioned challenges from different perspectives. On the one hand, to reduce extra computational costs, we fully leverage the features from different hidden layers for active sample selection and avoid applying many optimization techniques. On the other hand, instead of human experts (strong labeler) alone, we can hopefully engage machine experts (weak labeler) to provide reliable yet cheap annotations. Therefore, we propose a novel framework that combines AL and SSL, providing synergistic annotations from both strong and weak labelers.

To reduce the complexity of the algorithm and obtain satisfactory performance, we intend to exploit the knowledge within the networks themselves, minimizing extra computational time. It is well noted that the information in hidden layers of networks can be used to supervise the learning process [15]. Dou *et al.* [16] first introduces the deep supervision mechanism for liver segmentation, in which, the feature maps of some hidden layers in FCN are extracted to produce dense predictions for classification error estimation. Deeply supervised learning helps speed up the training process and improve the discrimination capability. It is observed that different hidden layers of networks have different performances on the same sample, while different samples obtain different feature maps from the same layer. We hypothesize that the dissimilarity between different feature maps can be used to examine uncertainty and provide guidance for AL, so we inject deep supervision into some hidden layers of U-Net [5] and design a novel AL strategy for biomedical image segmentation. The main contributions of the work are summarized as follows:

- We propose a novel deep active learning framework, DSAL, which combines AL and SSL, benefiting from both strong and weak labelers. The process of expert annotation is simulated using a fully annotated dataset, while an ensemble of denseCRF [17] serves as weak labelers to provide accurate pseudo labels.
- We propose a novel criterion for uncertainty evaluation, in which, multi-level segmentation results are integrated to form the basis for AL selection. In addition, a confidence criterion is designed to rectify the sample selection process and filter low-confidence samples.
- We evaluate DSAL on public biomedical datasets and compare it with a list of state-of-the-art AL methods. Experimental results show that DSAL significantly reduces the computational time and labeling costs without compromising the segmentation performance.

The remainder of the work is structured as follows. Section II reviews recent work on SSL and AL. The details of DSAL are described in Section III. The experimental datasets, setup, baselines and results are presented in Section IV and Section V. Finally, Section VI presents our conclusions.

II. RELATED WORK

Both semi-supervised learning and active learning aim to view the model optimization problem as a trade-off between two objectives: (i) Improve the performance and generalizability of the model among unseen data; (ii) Reduce the annotation costs in training data [9]. The former focuses on better leveraging unlabeled data, while the latter seeks to wisely select informative samples to meet a given annotation budget. In this section, we give a review of the relevant literature in semi-supervised learning and active learning.

A. Semi-Supervised Learning

SSL paradigm is always considered as a self-training or pseudo-labeling process, in which, the initialized model with a small labeled dataset is further trained with pseudo annotations of unlabeled data in an iterative manner [18]. Because no extra expert annotations are required, SSL has been popular in biomedical analysis. Bai *et al.* [3] propose a two-step iterative method for cardiac MR image segmentation, in which, a CNN trained on the labeled data is updated with pseudo annotations refined by the application of CRF. Similarly, Zhao *et al.* [1] propose a student-teacher framework which consists of a U-Net [5] for segmentation and a denseCRF for refinement. These methods are easy to implement but are limited by the quality of pseudo annotations. Therefore, the models can be improved by a narrow margin compared with labeled data only.

Min *et al.* [19] propose a two-stream mutual attention network to reduce the noise in pseudo labels. Xia *et al.* [20] suggest a semi-supervised framework embedded with ensemble learning. In the framework, labeled data and the corresponding ground truth are fed into multiple networks for co-training, and pseudo annotations are generated through the ensemble of network predictions. 5% improvement in Dice has been achieved over a full supervised model for pancreas segmentation. These works prove that unlabeled data and pseudo labels can be helpful for performance improvement when the labeled data is limited. Moreover, combining the decisions from multiple models can help improve the quality of pseudo annotations to some extent.

B. Active Learning

AL paradigm allows extra samples to be selected and labeled. Instead of using pseudo labels, AL interactively queries new instances annotated by the oracle (expert annotator) so that the performance of networks is boosted. Therefore, the key to AL becomes how to seek informative and valuable samples for labeling. Yang *et al.* [21] propose a two-stage framework named as Suggestive Annotation (SA), in which a set of Fully Convolutional Networks (FCNs) with initial labeled data are trained, and samples are selected to be labeled based on (1) the uncertainty identified through an ensemble of FCNs, and (2) the similarity between high-level features extracted from images. Then, a generalized maximum coverage algorithm is proposed to suggest samples for labeling. Ozdemir *et al.* [22] also propose to select samples based on uncertainty and representativeness. Instead of multiple models, Monte Carlo Dropout (MC dropout) [23] is proposed to evaluate the variance of outputs from the trained network by enabling dropout at inference time. However, these frameworks consider the samples equally, regardless of the imbalanced annotation costs, which lead to higher labeling costs in practice.

Kuo *et al.* [14] take annotation costs into account when selecting samples, and design an AL framework based on the
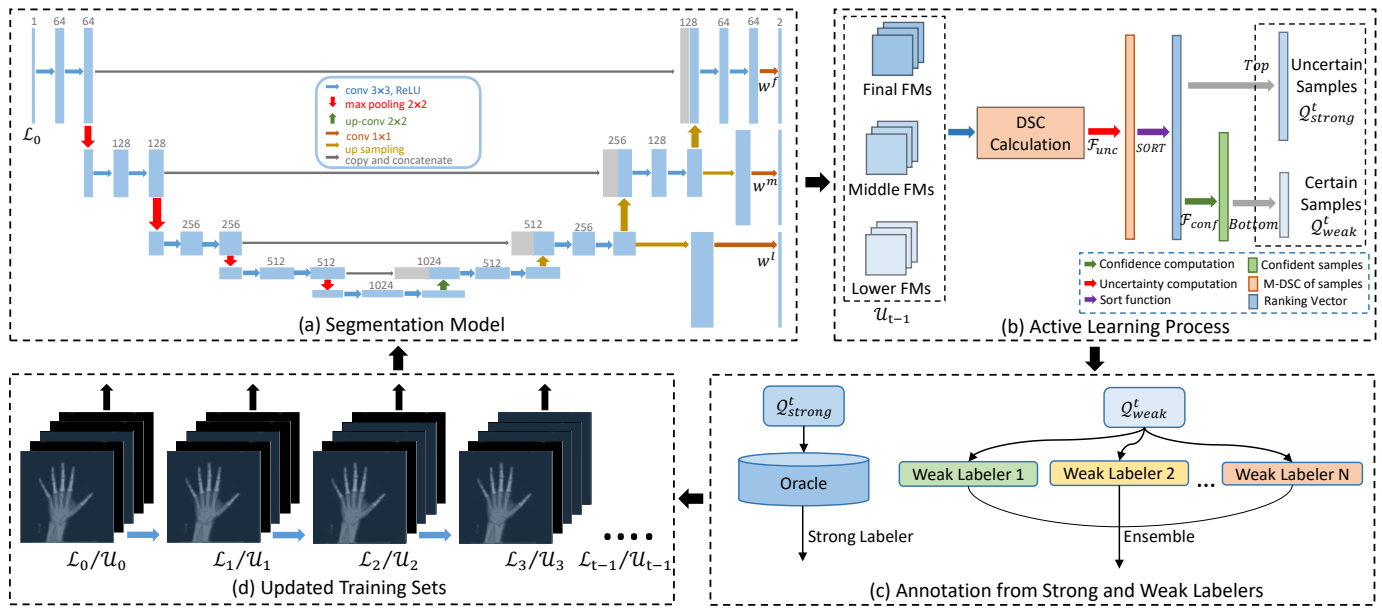
Fig. 1. Overall framework of our proposed method: (a) Deeply Supervised U-Net. (b) Active Learning Process. (c) Annotation from Strong and Weak Labelers. (d) Updated Training Sets.

uncertainty of multiple PatchFCNs, in which, labeling time is estimated based on mask boundary length and number of connected components using linear regression. They formulate a knapsack 0-1 problem for maximizing the uncertainty while maintaining the annotation costs. It is difficult to estimate labeling time for different modalities and tasks in biomedical imaging, for instance, segmentation on the left ventricle (LV) is more difficult than the whole heart. Another solution is to leverage pseudo labels for AL. Gorriz *et al.* [24] propose a cost-effective AL method based on MC dropout, or CEAL for short. In each iteration, highly uncertain samples are selected for annotation, while samples with low uncertainties along with their predicted masks are directly used in the next iterations. CEAL is evaluated on ISIC 2017 dataset [25] and achieved a Dice's Coefficient of 74% after 9 AL iterations.

Despite the difference in handling the labeling costs, these methods employ multiple models or variations of models to estimate the uncertainty of samples, which is slow and difficult to train, due to the repetitive training process. Besides, optimization techniques such as dynamic programming and greedy algorithm are applied in some AL methods, which further increase the computational complexity [26], [27].

III. PROPOSED METHOD

The proposed AL framework is illustrated in Fig. 1. U-Net with deep supervision (deeply supervised U-Net, DS U-Net) is adopted as the segmentation model, where multi-level deep supervision is injected into different hidden layers. First, the model is trained on a small amount of labeled data $\mathcal{L}_0$, In each AL iteration t, the predicted masks from different layers of the same segmentation model for the remaining unlabeled data $\mathcal{U}_{t-1}$ is gathered for the AL process. In the AL process, the quality of the samples can be estimated based on the confidence and uncertainty between the results of

different layers, and different samples with different qualities are dispatched to strong or weak labelers for annotation. In this way, the original training set is updated, and the segmentation model is fine-tuned with the updated dataset $\mathcal{L}_t$ incrementally.

A. Deeply Supervised U-Net

The segmentation model is shown in Fig. 1 (a). Different from earlier work [28], we adopt a light variant of U-Net [5], in which, upsampling layers only include nearest-neighbor interpolation instead of interpolation with 1×1 convolution. Upon the basic architecture of U-Net [5], we inject deep supervision into several hidden layers, namely *lower layer* and *middle layer*. Specifically, low-level and middle-level feature maps on the decoding stage are upscaled to original resolution using additional upsampling layers. Then, the softmax layer is applied to these layers to obtain predicted masks for final loss calculation. Let W be the weights of the U-Net, and w^l, w^m, w^f be the weights of three classifiers of *lower layer*, *middle layer* and *final layer*, respectively, then the cross-entropy loss function of one layer can be formatted as:

$$L(\mathcal{X}; W) = \sum_{x_i \in \mathcal{X}} -\log p\left(t_i \mid x_i; W\right), \quad (1)$$

where $\mathcal{X}$ denotes the training samples and $p\left(y_i = t\left(x_i\right) \mid x_i; W, w^c\right)$ is the probability of target class label $t\left(x_i\right)$ corresponding to sample $x_i \in \mathcal{X}$, in which $c \in \{l, m, f\}$ denotes the index of the classifiers. Finally, the total loss function can be defined as:

$$L\left(\mathcal{X}; W, w^l, w^m, w^f\right) = \sum_{c \in \{l,m,f\}} \alpha_c L_c\left(\mathcal{X}; W, w^c\right), \quad (2)$$

where $\alpha_l, \alpha_m, \alpha_f$ are the weights of the associated classifiers. To control the strength of different terms, in our experiments, α_l, α_m and α_f are set as 0.1, 0.3 and 0.6, empirically.

Finally, classifications at different levels are performed together, which helps the shallower layers to learn discriminative features more efficiently.

B. Active Learning Process

We define the AL process as follows: given a small labeled dataset $\mathcal{L}_0$ and unlabeled pool U_0, in each iteration t, find valuable samples from U_{t-1} for labeling and updating the model. Therefore, the core of AL is the criterion for informative sample selection. As mentioned in previous sections, our criterion utilizes the knowledge within the networks.

In DS U-Net, some hidden layers are supervised by optimizing the final loss function, and well-learned layer parameters always generate accurate predictions from the hidden layers. Moreover, we find that different hidden layers of the well-trained model provide consistent prediction masks for certain samples, the consistency of which can be used to evaluate the sample uncertainty. There, based on the influence of samples on the model, the predicted masks of the hidden layers (*lower layer* and *middle layer*) are extracted to calculate the Dice's Coefficient (DSC) [5] with the results of *final layer*, called *L-DSC* and *M-DSC*, respectively. Finally, the mean of *L-DSC* and *M-DSC* serves as a proxy of the quality for each sample, termed as *Mean-DSC*. Higher *Mean-DSC* means a smaller uncertainty score. The uncertainty criterion is defined as $\mathcal{F}_{unc}$, which output the data with *Mean-DSC*.

Once the uncertainty score *Mean-DSC* is defined, we can visualize its relation with the accuracy of the predictions. We explore the relationship between *real DSC (R-DSC)* and *Mean-DSC*, shown in Fig. 2. In the iterative training process, we randomly select 10 samples from each step and put them into the model for prediction, and then we sort the DSC by descending order in each iteration, and the scatter plots and trend lines between the *Mean-DSC* order and the *R-DSC* order are drawn. There is some strong relationship between these orders, which allows selecting two types of samples for labeling, as shown in the following.

(1) **Highly uncertain samples**: Candidates with high uncertainty to be annotated by the strong labelers (the oracle).

(2) **Certain samples**: The most confidently samples with high consistency, which will be annotated by the weak labelers (pseudo-labeling).
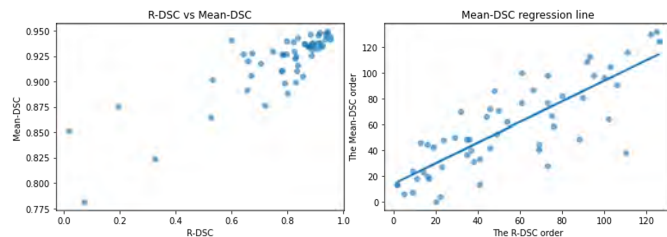

Fig. 2. The relationship between R-DSC and Mean-DSC on RSNA dataset. Left: the scatter plot based on (R-DSC, Mean-DSC). Right: DSCs are sorted by descending order. Linear regression was carried out with the scatters based on (R-DSC order, Mean-DSC order). Linear regression was carried out with the scatters (the coefficient is 0.80).

It is noted that low-variance results can be selected based on the aforementioned uncertainty ranking strategy among different hidden layers, but the models may have high biases on the predicted masks, which means the results with low confidence may be selected for labeling. The situation becomes noticeable especially on certain samples for pseudo-labeling, as in our definition they are with low prediction uncertainty. To guarantee the high segmentation quality of the selected samples for weak labelers, we calculated the confidence of the predicted segmentation mask by averaging the pixel-wise classification confidences, which are defined as:

$$\mathcal{C}(x_i) = \frac{1}{H \times W} \sum_{h=0}^{H} \sum_{w=0}^{W} \left\| \mathbf{P}_{hw} - \frac{1}{2} \right\|_1, \qquad (3)$$

where $\mathbf{P} \in \mathbb{R}^{H \times W}$ is the probability map of sample x_i. This definition presents how close the posteriors are to either background or foreground class. Therefore, the prediction confidence is effectively estimated when there is no available ground truth to be compared against. This confidence score serves as a complementary criterion to the uncertainty one which improves the selectivity of the latter.

As usual a threshold $t_{conf} \in (0,1)$ can be fixed to filter results with low confidence for a specific model architecture and dataset. However, in the AL scenario, different AL iterations have a wide range of confidence scores. Therefore, we design a histogram-based strategy for confidence implementation. In each AL iteration, a histogram with B bins can be generated by confidence scores of all remaining unlabeled samples, where only the uppermost bin is left for the following process, *i.e.*, t_{conf} is adaptively set to the beginning value of the uppermost bin. The confidence criterion is formulated as:

$$\mathcal{F}_{conf} : \mathcal{X} \rightarrow \{x_i \mid \mathcal{C}(x_i) > t_{conf}, x_i \in \mathcal{X}\}, \qquad (4)$$

where $t_{conf} = \frac{B-1}{B}(\max(\mathcal{C}(x_i)) - \min(\mathcal{C}(x_i))$. Fig. 3 visualizes the selection process involving both uncertainty and confidence criterion for certain samples.
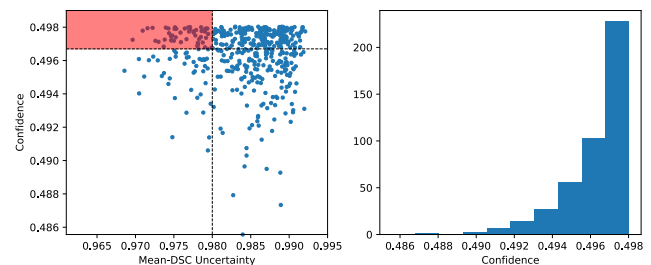

Fig. 3. Illustration of the active selection process when uncertainty and confidence are involved together. (Left) Confidence vs. Mean-DSC uncertainty and (Right) histogram of the confidence scores of query samples. First samples beyond the horizontal dashed line are selected according to the confidence range of the last bin in the corresponding histogram. Then the samples are ranked based on the uncertainty function $\mathcal{F}_{unc}$ and the samples to the left of the vertical dashed line are bottom-k ones. The red area delineates the most confidently samples to be annotated by weak labelers.

C. Annotation from Strong and Weak Labelers

In each AL iteration, the indices of highly uncertain samples and certain samples are exported for strong and weak labelers, respectively. We simulate the process of strong labelers using

the fully annotated dataset, namely $\mathcal{F}_{\text{strong}}$. For certain samples, normally, pseudo labels generated by DS U-Net can be used in a semi-supervised learning manner. To further refine these pseudo labels, denseCRF is utilized as a weak labeler.

In feature maps, every pixel i is a node with a label x_i and an observation value y_i, while the relationships among pixels are regarded as edges. The labels x_i behind pixels can be inferred by observations y_i, and the denseCRF I is characterized by a Gibbs distribution,

$$P\left(Y = y | I\right) = \frac{1}{Z(I)} \exp\left(-E(y|I)\right), \qquad (5)$$

where $E(y|I)$ is the Gibbs energy of a label y, which is formulated as

$$E(y) = \sum_i \Psi_u\left(y_i\right) + \sum_{i<j} \Psi_p(y_i, y_j), \qquad (6)$$

among which, the unary potential function $\Psi_u\left(y_i\right)$ is donated by the output of DS U-Net, and the pairwise potentials in our model is given by

$$\Psi_p(y_i, y_j) = \mu(y_i, y_j) \sum_{m=1}^{M} w^{(m)} k_G^{(m)}(f_i, f_j), \qquad (7)$$

where each $k_G^{(m)}$ is a Gaussian kernel $k_m(f_i, f_j)$, the vectors f_i and f_j are feature vectors for pixels i and j respectively, $w^{(m)}$ are linear combination weights, and μ is a label compatibility function. In practice, it is difficult to adjust these hyperparameters in denseCRF, which may lead to large variations in the refinement process. To reduce the performance variance, we propose to implement multiple denseCRFs in an ensemble learning manner to achieve robust generalization ability.

Specifically, we first search for a reasonably good base denseCRF specified by a set of hyperparameters including weights for both Gaussian and bilateral pairwise in Eq. 7, and the times of the iteration (step). The parameter search process is carried out using Tree-structured Parzen Estimator [29]. Next, to introduce ensemble variance, we randomly fluctuate each of the hyperparameters around their centers (see Tab. I for central hyperparameters configurations) by sampling a sharp normal distribution around the above denseCRF, forming a set of denseCRFs slightly different from each other but with relatively good performance. We then ensemble these denseCRFs together to obtain a strong learner achieved by element-wise majority voting. In case the ensemble learner still performs suboptimal compared to the average performance of the base learners, we adopt a greedy algorithm which can fine-tune the ensemble denseCRFs by repeatedly selecting the one with the largest dice score among the base learners as a new reference for parameter fluctuation and generate new strong learners. After several rounds of such fine-tuning processes, the denseCRF initialization procedure can be ended up with the optimized set of base learners, and then their parameters are kept constant and unchanged throughout the experiments.

The pseudocode of the proposed AL process is summarized in Algorithm 1. In each iteration of AL, the proposed algorithm selects both uncertain samples and certain samples from unlabeled dataset $\mathcal{U}_0$ based on confidence score and uncertainty score. After that, the segmentation model is fine-tuned with the enlarged dataset. Then the updated model is evaluated on an out-of-bag testing dataset. The process is repeated until exhausting the cost budget or reaching satisfactory performance. We can also set a fixed iteration times $\mathcal{T}$.

Algorithm 1: The proposed active learning process

Input : Small Labeled dataset $\mathcal{L}_0$, Unlabeled dataset $\mathcal{U}_0$, iteration times $\mathcal{T}$, uncertainty function $\mathcal{F}_{unc}$, confidence function $\mathcal{F}_{conf}$, query batch size $\mathcal{K}_{strong}$, $\mathcal{K}_{weak}$, masks generation functions $\mathcal{F}_{strong}$, $\mathcal{F}_{weak}$, base model $\mathcal{M}_0$.

Output: Labeled dataset $\mathcal{L}_t$, updated model $\mathcal{M}_t$.

1 **for** $t \leftarrow 1$ **to** $\mathcal{T}$ **do**

2 　　/* phase 1: query selection　　*/

3 　　$\mathcal{Q} \leftarrow \mathcal{F}_{unc}\left(\mathcal{U}_{t-1}, \mathcal{M}_{t-1}\right)$;

4 　　$\mathcal{Q}' \leftarrow \mathcal{F}_{conf}\left(\mathcal{Q}\right)$;

5 　　$\mathcal{Q}_{strong}^t \leftarrow \{\mathcal{U}_{t-1,i} \mid i \in \arg\underset{unc}{\text{topK}}\left(\mathcal{Q}; \mathcal{K}_{strong}\right)\}$;

6 　　$\mathcal{Q}_{weak}^t \leftarrow \{\mathcal{U}_{t-1,i} \mid i \in \arg\underset{unc}{\text{btmK}}\left(\mathcal{Q}'; \mathcal{K}_{weak}\right)\}$;

7 　　/* phase 2: sample annotation　　*/

8 　　$\tilde{\mathcal{Y}}_{strong}^t \leftarrow \mathcal{F}_{strong}\left(\mathcal{Q}_{strong}^t\right)$;

9 　　$\tilde{\mathcal{Y}}_{weak}^t \leftarrow \mathcal{F}_{weak}\left(\mathcal{Q}_{weak}^t\right)$;

10 　　/* phase 3: update model　　*/

11 　　$\mathcal{L}_t' \leftarrow \mathcal{L}_{t-1} \cup \{(\mathbf{x}, y) \mid \mathbf{x} \in \mathcal{Q}_{strong}^t, y \in \tilde{\mathcal{Y}}_{strong}^t\}$;

12 　　$\mathcal{L}_t \leftarrow \mathcal{L}_t' \cup \{(\mathbf{x}, y) \mid \mathbf{x} \in \mathcal{Q}_{weak}^t, y \in \tilde{\mathcal{Y}}_{weak}^t\}$;

13 　　$\mathcal{M}_t \leftarrow$ fine-tuning $\mathcal{M}_{t-1}$ using $\mathcal{L}_t$;

14 　　$\mathcal{U}_t \leftarrow \mathcal{U}_{t-1} \setminus \left(\mathcal{Q}_{strong}^t \cup \mathcal{Q}_{weak}^t\right)$;

15 **end for**

16 **return** $\mathcal{L}_{\mathcal{T}}, \mathcal{M}_{\mathcal{T}}$.

IV. Experiment Settings

Here, we first list and briefly describe the state-of-the-art baselines as follows:

- Random Query: randomly querying $\mathcal{K}_{strong}$ samples without replacement from the unlabeled pool $\mathcal{U}_0$.
- ensemble-S [21], [30]: an AL strategy based on uncertainty among an ensemble of T_e networks. In each iteration, the mean of Shannon entropies based on multiple networks is calculated as the uncertainty score for sample selection.
- ensemble-JS [14]: adopt similar uncertainty sampling with ensemble-S. The Jensen-Shannon divergence over the ensemble posteriors is computed.
- MC Dropout [22], [31]: Monte Carlo (MC) dropout is performed to obtain different class posterior probabilities in T_m MC parameter sets drawn from dropout distribution, which is used to measure uncertainty.
- CEAL [24]: a Cost-effective AL approach using dropout on unlabeled samples to estimate the pixel-wise uncertainty. Besides, unlabeled data with pseudo labels are selected for complementary sample selection.

A. Datasets

To demonstrate the effectiveness of DSAL, we carry out our experiments on two public biomedical datasets.

- ISIC 2017 dataset [25]: composed of 2000 RGB dermoscopy images for skin lesion analysis. Each image was annotated by medical experts with pixel-wise segmentation of the melanomas. Following the AL scenario in [24], 600 samples with annotations are randomly selected as $\mathcal{L}_0$, while 1000 images are randomly selected as Unlabeled pool $\mathcal{U}_0$ for iterative learning. The remaining 400 samples are used for testing.
- RSNA Bone Age dataset [7]: includes 12611 hand radiographs. We follow the preprocessing and sampling methods from [28] and obtained a small balanced dataset where $|\mathcal{L}_0| = 10$ and $|\mathcal{U}_0| = 129$. The evaluation is carried on a striped test set of 50 samples.

B. Technical Details

For the ISIC 2017 dataset, dermoscopic color images and corresponding binary masks are first converted into grayscale and normalized using z-score while also resized to 192×240 before training. We set $\mathcal{K}_{strong} = 35$ (import 35 oracle samples from the unlabeled pool $\mathcal{U}_0$ at the end of each AL iteration), while for both CEAL [24] and ours, pseudo-labeling process is involved from the 5th AL iteration onwards where we set both of $\mathcal{K}_{weak} = 20$ for fair comparison. We choose the dice loss as the objective function. During the training process, the base model is trained with a batch size of 32 for 15 epochs, and in each AL iteration, the model is fine-tuned for 10 epochs with an enlarged dataset. The learning rate is 1×10^{-5} throughout AL training.

For the RSNA Bone Age dataset, the images are resized into 512×512, followed by standardization with min-max normalization. During the data augmentation process, the images are randomly rotated by ± 0.2 degrees or flipped horizontally. We set $\mathcal{K}_{strong} = 10$ and $\mathcal{K}_{weak} = 10$ for each AL iteration. In CEAL [24], the proposed algorithm samples 3 images (No-detections), 3 images (Most uncertain), and 4 images (Random) for human annotations. We choose binary cross-entropy as the objective function. In each AL iteration, the DS U-Net is trained for 150 epochs with a batch size of 2, and the training process will be stopped if validation loss does not improve for 30 epochs. The initial learning rate is 1×10^{-4}, which will be reduced by an order of magnitude if validation loss does not improve for 30 epochs. We repeat the learning rate adjustment at most twice per iteration.

The experiments are conducted on an Intel Xeon Gold 6230 CPU and a Tesla V100 32GB GPU. The DS U-Net is optimized using Adam optimizer [32]. The selected denseCRF hyperparameters after adaptive searching are shown in Tab. I, while the number of bins in the confidence histogram for ISIC dataset and RSNA dataset are set to 10 and 5, respectively. For the sake of fairness, we implement the baselines by using the same segmentation architecture. In ensemble-S and ensemble-JS, an ensemble of $T_e = 3$ networks is created, while $T_m = 10$ MC parameter sets are drawn from dropout distribution.

TABLE I
CENTERS OF HYPERPARAMETERS OF DENSECRF FOR BOTH DATASETS

ISIC				
	sdims	schan	compat	step
Gaussian Pairwise	29.93	-	9.06	-
Bilateral Pairwise	28.19	5.59	9.46	-
Inference	-	-	-	2
RSNA				
	sdims	schan	compat	step
Gaussian Pairwise	1	-	6	-
Bilateral Pairwise	1	7	4	-
Inference	-	-	-	1

V. EXPERIMENTAL RESULTS

A. Study on RSNA Bone Age dataset

The experimental results on RSNA Bone Age dataset are shown in Fig. 4. We repeat the AL process for 12 iterations until all images are annotated. We can observe that the proposed method only uses 50.36% of annotations (6th iteration) to achieve comparable results as compared with full annotation, while other methods require more labeling resources. It is noted that the segmentation performance of ensemble-S and ensemble-JS is better than full annotation, since an ensemble of $T_e = 3$ networks always outperforms one alone. However, ensemble learning on networks is less effective and impractical for active learning, especially when T_e increases dramatically. The cost of labeling is studied on ISIC 2017 dataset, and discussed in Section V-B.
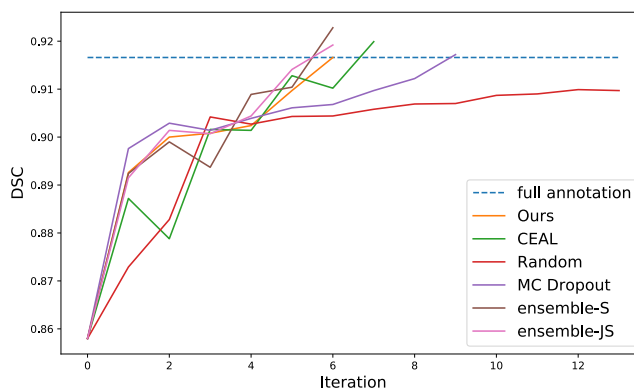

Fig. 4. Experimental results on RSNA Bone Age dataset.

B. Study on ISIC 2017 dataset

The experimental results on ISIC 2017 dataset are shown in Fig. 5 and Tab. II. We set a fixed iteration times $\mathcal{T} = 9$ as reported in CEAL [24]. The results show that the segmentation performance of the proposed method steadily increases with fewer fluctuations compared to other methods, especially in the last few iterations. It is noted that CEAL with DS U-Net reached Dice's Coefficient of 81.97 % after 9 AL iterations, which is better than 74 % reported in [24]. This demonstrates the effectiveness of the DS U-Net. Furthermore, our method achieves the best DSC after 5 iterations.
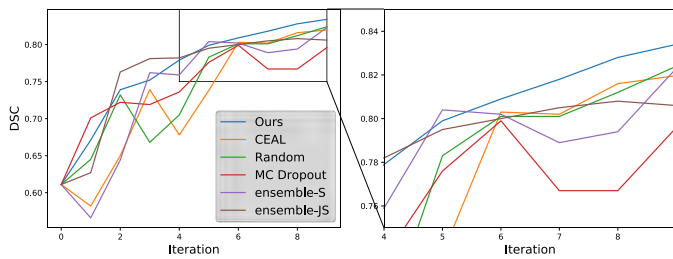
Fig. 5. Experimental results on ISIC 2017 dataset (details after 4 iterations are amplified on the right side).

We visualize some segmentation results on the testing dataset of different AL models in Fig. 6. In consistent with the performance in Fig. 5, ensemble-S can better capture the contours than ensemble-JS, yet both outperform the random query method. It is noted that the proposed method has less false positives compared to others, and generates more smooth contours. These benefit from the refinement of denseCRFs. Moreover, the involvement of multiple weak labelers can further improve the quality of pseudo labels. Besides, DS U-Net can properly grasp more meaningful and noiseless knowledge from different feature levels.

TABLE II
COMPARISON OF DICE'S COEFFICIENT ON ISIC 2017 DATASET AMONG DIFFERENT METHODS ACROSS AL ITERATIONS. MCD ABBREVIATES MC DROPOUT [22], [31], ENS-S ABBREVIATES ENSEMBLE-S [21], [30] AND ENS-JS ABBREVIATES ENSEMBLE-JS [14].

iter	Ours	Random	MCD	ens-S	ens-JS	CEAL [24]
base	0.611	0.611	0.611	0.611	0.611	0.611
1	0.671	0.645	**0.701**	0.566	0.627	0.582
2	0.739	0.732	0.722	0.644	**0.763**	0.650
3	0.752	0.668	0.719	0.762	**0.781**	0.739
4	0.779	0.705	0.736	0.759	**0.782**	0.678
5	0.799	0.783	0.776	**0.804**	0.795	0.738
6	**0.809**	0.801	0.799	0.802	0.800	0.803
7	**0.818**	0.801	0.767	0.789	0.805	0.802
8	**0.828**	0.812	0.767	0.794	0.808	0.816
9	**0.834**	0.824	0.796	0.823	0.806	0.820

Tab. III shows the average duration of all methods on the same machine. It is noted that the running time is subject to the device and its real-time workload. Given the same training strategy and networks, different active learning methods have similar training time, except ensemble-S and ensemble-JS, which require a large amount of repetitive training process, depending on T_e. These methods are much less practical in the real world where it could involve larger workloads on training and annotations, especially when T_e is increased. We noticed that our proposed method was faster than CEAL and MC dropout methods in the AL process, while labeling time for weak labelers (denseCRFs) is 7.59 seconds per sample, which is much less than manual labeling. The time interval in our method between two iterations is acceptable, which makes it possible to deploy the method on-the-fly, especially in the setting of IoMT and edge computing.

C. Ablation study

To evaluate the effectiveness of each component in our holistic framework, a series of ablation experiments are carried
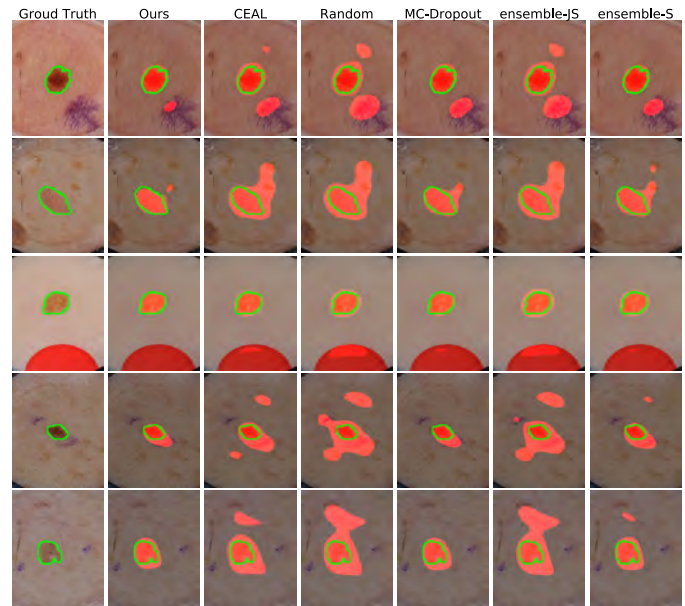

Fig. 6. Segmentation of **5** sample test images from ISIC 2017 dataset using different methods after the last AL iteration. The first column represents the original images layered with ground truth labels (bounded by green lines), and the rest show the segmentation results. The red pixels represent the segmentation masks by the corresponding methods. Ideally, pixels captured by green boundaries and red regions should be identical.

TABLE III
AVERAGE RUNNING TIME REQUIRED IN TRAINING AND AL QUERY FOR DIFFERENT METHODS ON ISIC 2017 DATASET.

	training (min/sample)	AL query (min/iteration)
Ours	**0.079**	**16.95**
CEAL [24]	0.079	88.74
MCD	0.079	95.58
ens-S	0.237	1.05
ens-JS	0.237	0.87
Random	0.079	1.05

out on ISIC dataset. As shown in Tab. IV, pseudo annotations can be regarded as complementary samples in the AL scheme. When the confidence criterion is involved, the segmentation performance is improved. Finally, the performance is further boosted by the refinement of ensemble denseCRFs, which demonstrates the effectiveness of multiple weak labelers. It is clear that each component of the proposed DSAL boosts the AL performance.

TABLE IV
ABLATION STUDY ON ISIC DATASET. DPL STANDS FOR SELECTING PSEUDO LABELS WITH M-DSC UNCERTAINTY MEASUREMENT AND ENSCRF STANDS FOR ENSEMBLE DENSECRF-BASED WEAK LABELING. DSC IS THE DICE SCORE OF SEGMENTATION ON TEST SAMPLES.

DS U-Net	dPL	Confidence	ensCRF	DSC
✓				0.8261
✓	✓			0.8276
✓	✓	✓		0.8321
✓	✓		✓	0.8334
✓	✓	✓	✓	0.8347

VI. CONCLUSIONS

In this work, we inject SSL into the AL scheme and propose a novel AL strategy for biomedical image segmentation. In the proposed framework, two types of annotations from diverse labelers are complementary to each other, and ensemble learning is applied to improve the performance of multiple weak labelers. A novel uncertainty measurement is proposed based on the disagreement of intermediate features in the deep supervision mechanism for active sample selection. In addition, a confidence-based criterion is adopted to rectify the sample selection process for pseudo-labeling. Experimental results demonstrate that the proposed AL strategy outperforms the baselines with the inclusion of human labeling costs and computational time.

REFERENCES

[1] Z. Zhao, X. Zhang, C. Chen, W. Li, S. Peng, J. Wang, X. Yang, L. Zhang, and Z. Zeng, "Semi-supervised self-taught deep learning for finger bones segmentation," in *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 2019, pp. 1–4.

[2] T. Han, V. X. Nunes, L. F. De Freitas Souza, A. G. Marques, I. C. L. Silva, M. A. A. F. Junior, J. Sun, and P. P. R. Filho, "Internet of medical things—based on deep learning techniques for segmentation of lung and stroke regions in ct scans," *IEEE Access*, vol. 8, pp. 71 117–71 135, 2020.

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

[4] N. Scarpato, A. Pieroni, L. Di Nunzio, and F. Fallucchi, "E-health-iot universe: A review," *management*, vol. 21, no. 44, p. 46, 2017.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[6] Y. Wu, K. Shen, Z. Chen, and J. Wu, "Automatic measurement of fetal cavum septum pellucidum from ultrasound images using deep attention network," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 2511–2515.

[7] D. B. Larson, M. C. Chen, M. P. Lungren, S. S. Halabi, N. V. Stence, and C. P. Langlotz, "Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs," *Radiology*, vol. 287, no. 1, pp. 313–322, 2017.

[8] Z. Zeng, X. Yang, Y. Qiyun, Y. Meng, and Z. Le, "Sese-net: Self-supervised deep learning for segmentation," *Pattern Recognition Letters*, vol. 128, 08 2019.

[9] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical Image Analysis*, p. 101693, 2020.

[10] V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Medical image analysis*, vol. 54, pp. 280–296, 2019.

[11] S. Dutt Jain and K. Grauman, "Active image segmentation propagation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2864–2873.

[12] R. S. Bressan, G. Camargo, P. H. Bugatti, and P. T. M. Saito, "Exploring active learning based on representativeness and uncertainty for biomedical data classification," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 6, pp. 2238–2244, 2018.

[13] Z. Zhou, J. Shin, L. Zhang, S. Gurudu, M. Gotway, and J. Liang, "Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7340–7351.

[14] W. Kuo, C. Häne, E. Yuh, P. Mukherjee, and J. Malik, "Cost-sensitive active learning for intracranial hemorrhage detection," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 715–723.

[15] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "-supervised nets," in *Artificial intelligence and statistics*, 2015, pp. 562–570.

[16] Q. Dou, L. Yu, H. Chen, Y. Jin, X. Yang, J. Qin, and P.-A. Heng, "3d deeply supervised network for automated segmentation of volumetric medical images," *Medical image analysis*, vol. 41, pp. 40–54, 2017.

[17] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in neural information processing systems*, 2011, pp. 109–117.

[18] H. Wu and S. Prasad, "Semi-supervised deep learning using pseudo labels for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1259–1270, 2017.

[19] S. Min, X. Chen, Z.-J. Zha, F. Wu, and Y. Zhang, "A two-stream mutual attention network for semi-supervised biomedical segmentation with noisy labels," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 4578–4585.

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

[21] L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen, "Suggestive annotation: A deep active learning framework for biomedical image segmentation," in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, Eds. Cham: Springer International Publishing, 2017, pp. 399–407.

[22] F. Ozdemir, Z. Peng, C. Tanner, P. Fuernstahl, and O. Goksel, "Active learning for segmentation by optimizing content information for maximal entropy," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, D. Stoyanov, Z. Taylor, G. Carneiro, T. Syeda-Mahmood, A. Martel, L. Maier-Hein, J. M. R. Tavares, A. Bradley, J. P. Papa, V. Belagiannis, J. C. Nascimento, Z. Lu, S. Conjeti, M. Moradi, H. Greenspan, and A. Madabhushi, Eds. Cham: Springer International Publishing, 2018, pp. 183–191.

[23] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, 2016, pp. 1050–1059.

[24] M. Górriz, X. Giró Nieto, A. Carlier, and E. Faure, "Cost-effective active learning for melanoma segmentation," in *ML4H: Machine Learning for Health NIPS, Workshop at NIPS 2017*, 2017, pp. 1–5.

[25] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in *2018 IEEE 15th Int. Symp. Biomed. Imag. (ISBI 2018)*. IEEE, 2018, pp. 168–172.

[26] D. S. Hochbaum, "Approximating covering and packing problems: set cover, vertex cover, independent set, and related problems," in *Approximation algorithms for NP-hard problems*, 1996, pp. 94–143.

[27] U. Feige, "A threshold of ln n for approximating set cover," *Journal of the ACM (JACM)*, vol. 45, no. 4, pp. 634–652, 1998.

[28] Z. Zhao, X. Yang, B. Veeravalli, and Z. Zeng, "Deeply supervised active learning for finger bones segmentation," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 2020, pp. 1620–1623.

[29] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Advances in neural information processing systems*, 2011, pp. 2546–2554.

[30] W. H. Beluch, T. Genewein, A. Nürnberger, and J. M. Köhler, "The power of ensembles for active learning in image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9368–9377.

[31] Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian active learning with image data," in *PMLR*, ser. Proceedings of Machine Learning Research, vol. 70, 2017, pp. 1183–1192.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.