Leveraging Large-Scale Weakly Labeled Data for Semi-Supervised Mass Detection in Mammograms

Yuxing Tang[1], Zhenjie Cao[1], Yanbo Zhang[1], Zhicheng Yang[1], Zongcheng Ji[1],
Yiwei Wang[1], Mei Han[1], Jie Ma[2], Jing Xiao[3], Peng Chang[1]
[1]PAII Inc., USA [2]Shenzhen People's Hospital, China [3]Ping An Technology, China
tangyuxing87@gmail.com, pengchang@gmail.com

Abstract

Mammographic mass detection is an integral part of a computer-aided diagnosis system. Annotating a large number of mammograms at pixel-level in order to train a mass detection model in a fully supervised fashion is costly and time-consuming. This paper presents a novel self-training framework for semi-supervised mass detection with soft image-level labels generated from diagnosis reports by Mammo-RoBERTa, a RoBERTa-based natural language processing model fine-tuned on the fully labeled data and associated mammography reports. Starting with a fully supervised model trained on the data with pixel-level masks, the proposed framework iteratively refines the model itself using the entire weakly labeled data (image-level soft label) in a self-training fashion. A novel sample selection strategy is proposed to identify those most informative samples for each iteration, based on the current model output and the soft labels of the weakly labeled data. A soft cross-entropy loss and a soft focal loss are also designed to serve as the image-level and pixel-level classification loss respectively. Our experiment results show that the proposed semi-supervised framework can improve the mass detection accuracy on top of the supervised baseline, and outperforms the previous state-of-the-art semi-supervised approaches with weakly labeled data, in some cases by a large margin.

1. Introduction

Mammography is widely used for the early detection of breast cancer, the most common cancer in women. With about 48 million mammograms performed annually in the US, a Computer-Aided Diagnosis (CAD) system for mammogram screening is of great interest. This paper focuses on detecting mass lesions from mammograms, an important indication of potential malignancy. Despite the recent progress of Deep Neural Networks (DNN)-based approaches, mass detection with high accuracy still remains a challenging problem.

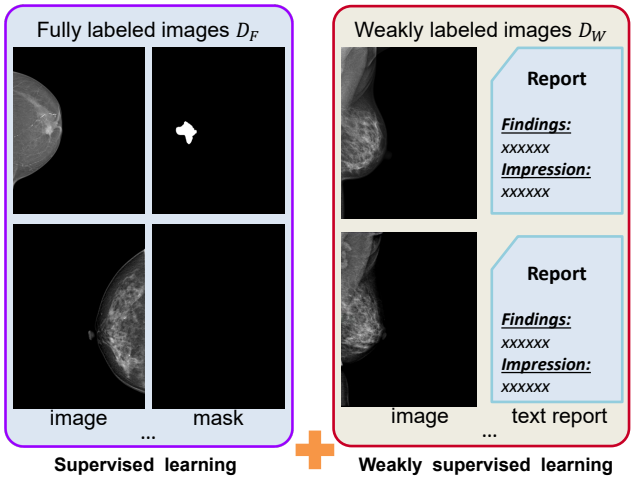

Figure 1. Our semi-supervised breast mass detection system uses fully labeled data (with mass masks) as well as large-scale weakly labeled data (with probabilistic NLP labels).

Many previous DNN-based mass detection models are trained with fully supervised fashion [26, 36, 22], and in turn require a large amount of fully labeled data (with pixel-level masks or rectangular bounding-boxes), which must be manually annotated by medical experts [33]. In many cases, the amount of high-quality annotation data has become the bottleneck for further performance improvement. Semi-supervised approaches can minimize the effort required to prepare the labeled data by training the model with a limited number of fully labeled examples and an additional, arbitrary amount of unlabeled or weakly labeled (*e.g.*, image-level class labels) examples [28, 30].

For this study, we collected a large-scale dataset containing 134,520 images with diagnostic reports from 30,495 patients, among which 2,634 images have pixel-level mass masks labeled by our collaborative radiologists. To fully leverage the diagnostic reports, we further adopt a *RoBERTa* [21]-based natural language processing (NLP)

model *Mammo-RoBERTa*, and it can take a diagnostic report as input and generate the probability of mass presence for the corresponding image, which is subsequently assigned as the image-level *soft* label.

To fully exploit the large-scale weakly labeled data (with image-level soft labels), and small amount fully labeled data (with pixel-level mass masks), we present a novel high-resolution, multi-stage semi-supervised learning framework, inspired by the previous self-training-based semi-supervised approaches [11, 24] and Multiple Instance Learning (MIL)-based weakly supervised learning methods [18, 1, 4, 29]. Starting from a high-resolution bilateral dual-view-based mass detection model first trained on the fully labeled data, the proposed framework iteratively updates the detection model in a self-training fashion, by carefully designing a novel batch sample mining strategy to identify the most informative samples for the next iteration. Compared to previous MIL-based approaches, a *soft* cross-entropy loss and a *soft* focal loss are deployed for the image-level classification task and the pixel-level classification task respectively, which are shown to outperform the hard label-based counterpart by our experiment results.

Meaningful performance improvement of the mass detection model has been achieved by applying the proposed framework to our large-scale weakly labeled dataset, in comparison with the same model trained with only fully labeled data. The proposed framework also outperforms various state-of-the-art semi-supervised 2D medical image detection and segmentation approaches, when the entire dataset is used. It is also worth noting that the model performance is monotonically increasing with respect to the amount of the weakly labeled data available in the training set (Section 4.4 Figure 4), while some previous work [18] may observe degraded model performance once the weakly labeled data exceeds a certain amount.

The main thrusts of this paper are summarized as follows.

1. We propose a novel self-training-based semi-supervised learning framework, by leveraging probabilistic soft class labels generated from the diagnostic report by an NLP model (*Mammo-RoBERTa*).

2. We propose a novel sample mining strategy to select the most informative weakly labeled data for each iteration of self-training, along with the *soft* label-based image-level and pixel-level classification loss functions.

3. We show meaningful improvement of the semi-supervised approach when compared with the previous fully supervised approaches, in terms of mass detection accuracy. In addition, the proposed framework can outperform several previously reported state-of-the-art semi-supervised approaches on our entire large-scale dataset.

2. Related Work

Weakly/Semi-supervised learning for object detection or segmentation. Semi-supervised learning approaches have been explored due to the lack of strongly labeled examples with object-level or pixel-level annotations. In specific, Expectation-Maximization (EM), consistency regularization [13, 17, 6] and pre-training [7] strategies have been applied to utilize a large amount of unlabeled data. In particular, EM-based self-training methods [24, 34, 11] treat the object locations or pixel-level masks as latent variables and alternately estimate the latent labels and optimize the DNN parameters in an iterative way. We also adopt the iterative self-training scheme in this work.

To utilize the image-level weakly labeled data, the multiple instance learning technique is applied in the medical image domain by enforcing at least one image patch belonging to the image labeled disease [18]. Several methods [23, 19] introduce an additional branch to perform image-level classification. A model is jointly trained for segmentation/detection and classification tasks to exploit the information contained in weakly labeled images, therefore the backbone convolutional layers are optimized with both image-level and pixel-level images together. The aforementioned methods benefit from a large amount of weakly labeled data for the case with only a very small group of fully-labeled data (typically from tens of to around a hundred images). However, the merit of weakly labeled data is remarkably compromised with an increased amount of fully-labeled data. In comparison, our developed self-training based method can constantly achieve outstanding performance with even thousands of fully labeled mammograms, which is more practical for developing a commercial computer-aided diagnostic system.

Many methods have been proposed to generate labels in various manners. Work in [3] estimates pseudo pixel-level labels for unlabelled data. Active learning allows a user to interactively label some new representative samples [35, 16], however, it is still expensive for medical applications. On the contrary, NLP techniques have been used to generate image labels from radiological reports [32, 5] to build a large scale medical image dataset. However, it contains some false labels due to the ambiguous description in reports. Unlike the aforementioned approaches, we generate soft labels from radiological reports using NLP to obtain a large scale image-level labeled data with minor bias. Moreover, we leverage a similar strategy as [3] to produce more accurate pseudo pixel-level labels with the help of soft-label information.

Lesion detection in mammogram. A standard mammography screening procedure acquires x-ray images from two projection views for each breast, respectively referred to as the craniocaudal (CC) view and the mediolateral oblique (MLO) view. Recently, some supervised learning-

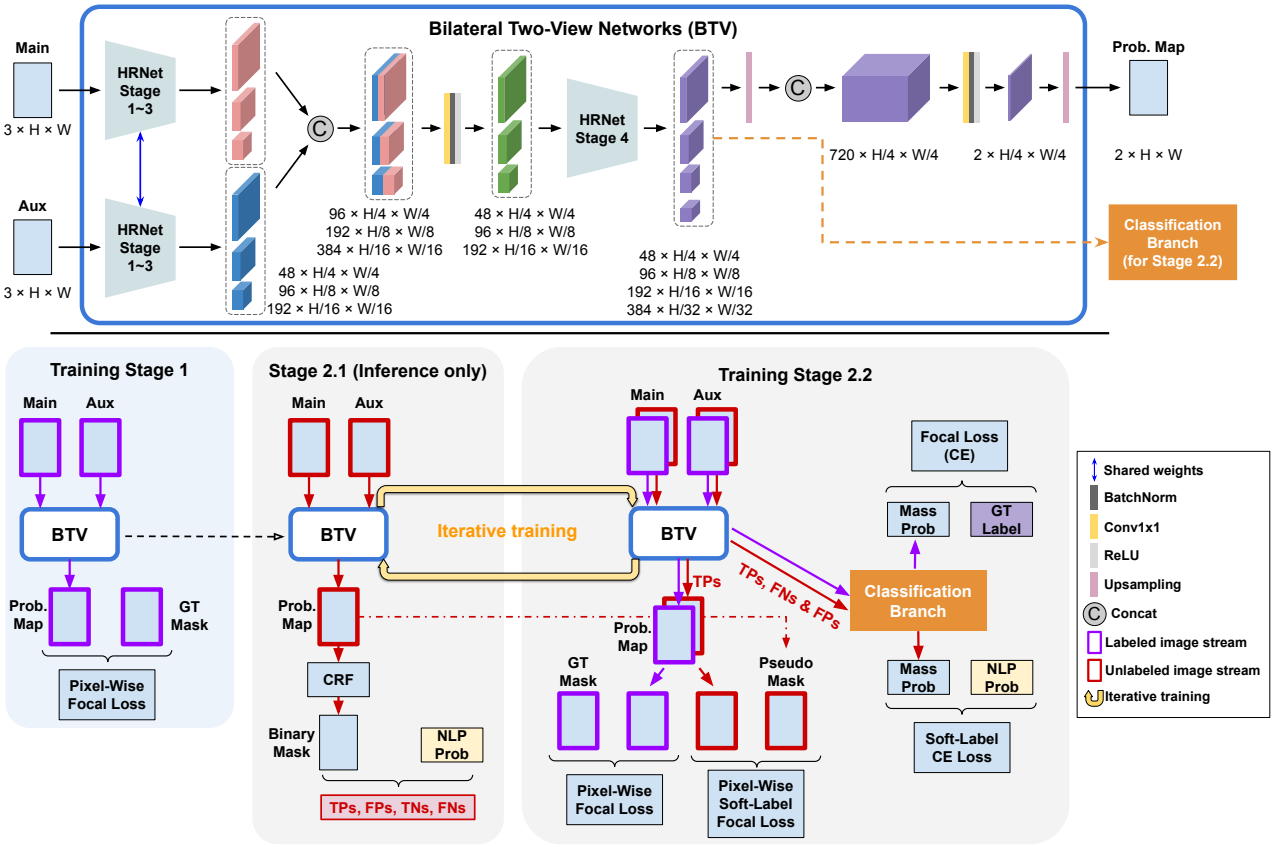

Figure 2. Overview of the proposed semi-supervised learning approach. The framework consists of a supervised learning model (Stage 1) with fully labeled images and an alternate and iterative training procedure (Stage 2.1 and 2.2) using a mixture of fully labeled images and weakly labeled images by the NLP technique. We design a high-resolution Bilateral Two-View Network for the mass detection task.

based methods have been proposed to detect mass from single-view [26] or multi-view mammograms [36, 22]. The later manner achieves superior performance with the assistance of mutual information from cross-view. Therefore, we also conduct detection on multi-view mammograms in this study. Weakly supervised learning approaches have been applied to detect lesions in mammograms using only image-level labeled data [1, 19, 4, 29]. Most of them consist of image classification and lesion localization branches using a similar strategy as MIL.

3. Proposed Method

We propose a novel semi-supervised learning framework to leverage a small number of fully labeled and numerous weakly labeled images (with diagnostic reports) for medical image detection or segmentation, and apply it for breast mass detection [1] in mammograms.

[1] While detection and segmentation are different computer vision tasks, we use segmentation and detection interchangeably in this paper. In fact, our framework is independent of segmentation or detection backbones. We design a breast mass segmentation network and evaluate its performance using the widely-used and clinically accepted mass detection metrics.

Mathematically, given a fully labeled dataset $\mathcal{D}_{\mathcal{F}} = \{(x_m, y_m), m = 1, 2, \cdots, M\}$ and a weakly labeled set $\mathcal{D}_{\mathcal{W}} = \{(x_n, y_n), n = 1, 2, \cdots, N\}$, where $x \in \mathbb{R}^{3\times H\times W}$ is an image and y is its label. $y_m \in \{0, 1\}^{H\times W}$ is manually annotated at pixel-level by medical experts, while y_n is initially unknown but can be later extracted from the diagnostic report at image-level (*i.e.*, $y_n \in \{0, 1\}^{1\times 1}$). In our case, the size of labeled dataset is much smaller than the weakly labeled set ($M \ll N$). The goal of this study is to estimate the parameters Θ of the detection/segmentation network and update the predictions on weakly labeled data alternately, so as to accurately predict pixel-wise label y_k from testing image x_k in $\mathcal{D}_{\mathcal{T}} = \{(x_k, y_k), k = 1, 2, \cdots, K\}$.

Figure 2 depicts the pipeline of our proposed framework. A supervised framework is first trained using pixel-level mask annotations (Training Stage 1). To leverage a vast number of images with only text reports, we develop a natural language processing (NLP) model to predict the probability of the presence of breast mass in each image/breast. We then iteratively mine these data using the probabilistic labels (Stage 2.1) and alternately training the network (Training Stage 2.2) with both sources of data using novel

sample mining strategy and soft-label loss functions.

3.1. High-resolution bilateral two-view networks

The first stage of the framework is a supervised pixel-level classification model trained on $\mathcal{D}_{\mathcal{F}}$. It takes a small number of high-resolution mammogram images and lesion masks annotated by radiologists as input. We design a deep high-resolution framework called *Bilateral Two-View Networks* (BTV) for this stage. The BTV is based on the HRNet [31] for semantic segmentation. We modify it to take bilateral mammograms as input (*e.g.*, RCC as the main image and LCC as the auxiliary image, or LMLO as the main image and RMLO as the auxiliary image, etc., where "L" and "R" respectively represent left/right breast.)

The fully supervised learning framework is formulated to optimize model parameters Θ:

$$\min_{\Theta} \sum_{m=1}^{M} L_{\text{seg}}(f(x_n;\Theta), y_m), \tag{1}$$

where $f(\cdot)$ is the pixel-level classification network parameterized by Θ and L_{seg} is the supervised loss function.

3.2. NLP labeling from text reports

Fully labeled medical images are usually scarce to train a supervised model, while a substantial number of medical images and their associated diagnosis reports are stored in many hospitals' Picture Archiving and Communication Systems (PACS). To take full advantage of these unlabeled data (or weakly labeled data considering that they have clinical reports), we develop a clinical natural language processing (NLP) model called *Mammo-RoBERTa* to predict whether or not a patient has breast mass given her mammography report on a large-scale dataset $\mathcal{D}_{\mathcal{W}}$ from the PACS. We formulate this problem as a text classification problem and propose a binary classification model by fine-tuning the pre-trained RoBERTa [21] with whole word masking in Chinese text [8]. We build separate prediction models for each (left and right) side of the breast.

As shown in Figure 3, the language classification model treats each text report as a sample and takes both the *description* and *impression* parts of a report as the input. Specifically, for each *description* $\mathbf{w}^{des}$ and *impression* $\mathbf{w}^{imp}$, we construct a sequence [CLS] $\mathbf{w}^{des}$ [SEP] $\mathbf{w}^{imp}$, where [CLS] is the special token used for the classification output, and [SEP] is the special token used for concatenating $\mathbf{w}^{des}$ and $\mathbf{w}^{imp}$. Suppose $\mathbf{w}^{des}$ and $\mathbf{w}^{imp}$ has n and m words, respectively, the input sequence is formulated as: [CLS], w_1^{des}, ..., w_n^{des}, [SEP], w_1^{imp}, ..., w_m^{imp}.

The pre-trained RoBERTa model with whole word masking in text [8] consists of 12 encoding layers with H=768 hidden units and 12 attention heads. We use the final hidden vector $C \in \mathbb{R}^{1\times H}$ corresponding to the first input token [CLS] as the aggregate representation of the input sequence. We apply a linear classification layer with weights

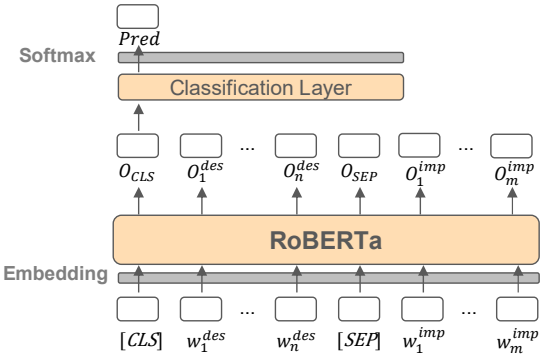

Figure 3. The architecture of the mammography report classification model: *Mammo-RoBERTa*.

$W \in \mathbb{R}^{2\times H}$ on C and use a softmax function to obtain the probability of an image x_i containing mass, formulated as:

$$\hat{y}_i = \text{softmax}(CW^T). \tag{2}$$

During training, since our labeled data is limited, we use three-fold cross-validation to make full use of the labeled data $\mathcal{D}_{\mathcal{F}}$, and at the same time to obtain more robust results. The goal is to minimize the cross-entropy loss over all the training examples:

$$L_{\text{NLP}} = \frac{1}{T}\sum_{i=1}^{T}\left[-y_i \log \hat{y}_i - (1-y_i)\log(1-\hat{y}_i)\right], \tag{3}$$

where T is the number of training data, $y_i \in \{0,1\}$ and $\hat{y}_i \in [0,1]$ are the ground-truth label and predicted probability of an image x_i containing breast mass, respectively.

After training, we obtain three models, each corresponding to one pair of two folds as the training data and one fold as the hold-out data. At test time, we apply these three models on the dataset $\mathcal{D}_{\mathcal{W}}$, and compute the average of the probabilities returned by the three models as the final probability y_n of an image x_n containing breast mass: $y_n = p(\text{mass}|x_n) \in [0,1]$. The unlabeled images thus become weakly labeled ones for our mass detection task by leveraging the NLP.

3.3. Iterative mining the weakly labeled data

To effectively leverage the large-scale weakly labeled data and probabilistic labels extracted using our Mammo-RoBERTa, we iteratively mine the data and feed the most *informative* samples to the framework for the next training stage, to improve the current BTV model. The informative samples include the data: 1). where the current image model possibly gets contradictory or inconsistent predictions with the language model; 2). which are likely to augment the current labeled dataset with potentially accurate pseudo masks. To obtain these informative samples, we design the following iterative mining strategy:

1. For each of the weakly labeled images x_n in $\mathcal{D}_{\mathcal{W}}$, we apply the trained BTV model in Stage 1 to generate a mass probability map (or prediction map) $M_p(x_n)$ and a

	NLP (+)	NLP (−)
Mask prediction (+)	TP	FP
Mask prediction (−)	FN	TN

Table 1. The definition of *tentative* TP (true positive), FP (false positive), FN (false negative), and TN (true negative) samples. Take the FP as an example, it is defined as an image that the NLP model predicts with a low probability (*i.e.*, tentative negative mass image), but the binary mask prediction indicates a mass detection (*i.e.*, positive mass detection).

binary mass prediction mask $M_b(x_n)$ by post-processing the probability map using a Dense Conditional Random Field (Dense CRF) [15].

2. After being labeled by Mammo-RoBERTa, each image x_n in $\mathcal{D}_{\mathcal{W}}$ gets a probability y_n, indicating the confidence of mass present within this image (or report, literally). For images with very high and low probabilities ($y_n > p_0$ or $y_n < 1 - p_0$), we tentatively treat them as positive and negative images, respectively, where the language model is confident with its predictions. We then compare the NLP labels and mass prediction masks at the current stage to mine a smaller number of informative samples from $\mathcal{D}_{\mathcal{W}}$. We define four kinds of *tentative* samples $\{\text{TPs, FPs, TNs, FNs}\} \in \mathcal{D}_{\mathcal{W}}$ in Table 1.
3. Since the majority of images are free of mass (healthy patient accounts for a substantial part of the screening mammography) in $\mathcal{D}_{\mathcal{W}}$, the true negatives (TNs) dominate the sample set. However, these TNs are less informative because the current BTV network is already capable of recognizing easy negatives. Consequently, we discard these tentative TNs and only feed the other three types of samples, which are more informative (large entropy from information theory), to the network for iterative re-training. We add an image-level classification branch (mass vs. no mass) to the BTV model and fine-tune the whole model with both samples from $\mathcal{D}_{\mathcal{F}}$ and $\mathcal{D}_{\mathcal{W}}$ (Stage 2.2 in Figure 2). To go into greater detail,
   - The **fully-labeled images** in $\mathcal{D}_{\mathcal{F}}$ go through the segmentation branch of the BTV model and the newly-added classification branch for image-level classification. The pixel-wise annotations labeled by radiologists and binary image-level labels indicating the presence or absence of breast mass are served as ground-truth. We adopt pixel-wise focal loss for the segmentation branch ($L^{\text{F}}_{\text{seg}}$) and binary cross-entropy loss for the classification branch ($L^{\text{F}}_{\text{cls}}$). Adding $\mathcal{D}_{\mathcal{F}}$ in the iterative training stage helps not only training the newly-added classification branch but also avoiding the BTV to be dominated by weakly labeled data during re-training.
   - The **weakly labeled images** in $\mathcal{D}_{\mathcal{W}}$: **TP**s go through both the segmentation and classification branches. Since their ground-truth image and pixel level labels are unknown, we use pseudo ground-truths, namely, pixel-wise probabilities from Stage 2.1 as segmentation mask, and NLP probabilities as classification labels. We design a pixel-wise soft-label focal loss and a soft-label cross-entropy loss for segmentation ($L^{\text{W}}_{\text{seg}}$) and classification ($L^{\text{W}}_{\text{cls}}$) based on the pseudo probabilistic labels, respectively (Please refer to Section 3.4 for our loss functions). **FN**s and **FP**s go through the shared DNN backbone (BTV) but only flow into the classification branch ($L^{\text{W}}_{\text{cls}}$) because the pseudo masks are either empty or could not provide any correct supervision to the segmentation branch. We use the NLP probabilities as ground-truth image labels and soft-target cross-entropy loss for FNs and FPs.
4. Update the BTV model with the latest model and iteratively perform step 1-3 Q times until converging (or performance saturated). Note that the TPs, TNs, FPs, and FNs are tentative and interchangeable with regard to the output of the current model during training. Ideally, the number of TPs and FPs will decrease along the iterative training process. We decrease p_0 within each training iteration to allow more weakly-labeled data. Please refer to implementation details in Section 4.2.

3.4. Pseudo labels and soft-label losses

Instead of direct using all the pseudo-binary masks predicted by BTV in Stage 2.2 and the binary NLP labels to alternately train the framework, we design two soft-target loss functions to better capture the probabilistic information. Using binary hard-labels generated by tuning the threshold is laborious and might be error-prone since 1). the prediction maps of supervised BTV may be inaccurate when trained with only a few data, 2). radiologist sometimes uses vague descriptions when he/she is uncertain about the presence of lesions and refers to further imaging or biopsy investigation. A soft probabilistic label from the NLP model implicitly indicates the confidence of the radiologist's interpretation. Moreover, training with inaccurate hard-labels might probably cause the network to get stuck in fake 'local minimum'. The soft-label strategy also enables the semi-supervised network to train with the NLP labeled data gradually based on the confidence of the NLP model – generally, the larger the confidence score (probability), the more accurate the label it be. The semi-supervised network selects images with more accurate labels first and gradually accepts noisy labels when it is more robust.

The soft-label cross-entropy loss is similar to the standard cross-entropy loss:

$$L_{\text{CE}} = \frac{1}{N}\sum_{n=1}^{N}\left[-y_n \log \hat{y_n} - (1-y_n)\log(1-\hat{y_n})\right], \quad (4)$$

where we replace the hard-label $y_n \in \{0, 1\}$ with soft probabilistic label $y_n \in [0, 1]$ as the ground-truth label, and $\hat{y_n}$ is the prediction (probability) of the network. This loss can be also computed at pixel-level for segmentation (*i.e.*, pixel-level classification).

Since there are far more negative data than positive data both at image-level and pixel-level, we adopt Focal loss [20] to address the issue of the class-imbalance problem. The Focal loss also considers the weights of easy and hard examples. We modify it to be compatible with probabilistic labels y_n. The soft-label focal loss is defined as:

$$L_{\text{S-FL}} = \alpha_t \cdot (y_n - \hat{y}_n)^{\gamma} \cdot L_{\text{CE}}, \tag{5}$$

where α_t and γ are pre-defined hyper-parameters used to balance the class weights and difficulties, respectively.

The overall loss of the semi-supervised model is:

$$L = L^{\text{F}}_{\text{seg}} + \lambda_1 L^{\text{F}}_{\text{cls}} + \lambda_2 L^{\text{W}}_{\text{seg}} + \lambda_3 L^{\text{W}}_{\text{cls}}. \tag{6}$$

4. Experiments

4.1. Data

We collected a large-scale database of 4-view mammograms and reports from our collaborating hospitals' PACS for several consecutive years, containing 134,520 images and over 33,630 reports [2]. Among these data, we have 2,634 images delineated with pixel-level mass masks by radiologists and 14,378 images free of mass according to radiologist's interpretation, *i.e.*, 3,753 labeled studies in total (each study has two images for each side of the breast: CC and MLO view, and any associated reports). In addition to the labeled imaging studies, we have 28,688 imaging studies (114,752 images) with mammography reports in Chinese. The labeled images were split into training and validation randomly at the patient-level with a ratio of 10:1. To evaluate the performance of the proposed model and to compare it with other methods, we have a hold-out testing set of 689 studies (2,756 images) manually labeled by radiologists.

4.2. Training

Networks. For the Bilateral Two-View (BTV) networks, we modify the HRNet-W48 for semantic segmentation [31] to fit the high-resolution, bilateral view mammograms. More specifically, we concatenate the three-resolution feature maps of the main image (*e.g.*, LCC view) and its auxiliary image (*e.g.*, RCC view) produced by the stage 1-3 of the HRNet, followed by 1×1 convolutions and non-linear operations (see the top figure of Figure 2). The structure of the newly-included classification branch is similar to that of HRNet-W48 for ImageNet classification, with 2 output units (instead of 1,000 for ImageNet classification) to predict the presence or absence of the mass. This classification branch is connected to stage 4 of the HRNet, which downsamples the four-resolution feature maps into small resolution and then aggregates them to obtain a final representation for classification. The proposed framework is independent of the DNN backbone (*e.g.*, ResNet [10], DenseNet [12], or U-Net [27]) and architecture (*e.g.*, Faster R-CNN [25] or Mask R-CNN [9]). We adopt the HRNet since it can maintain high-resolution representations through the whole process, which are crucial for our position-sensitive mass detection task.

[2]Institutional Review Board (IRB) number: LL-XJS-2020011

	Average F1	Weighted F1	Accuracy
left	0.9729	0.9873	0.9872
right	0.9581	0.9819	0.9819

Table 2. Results of three-fold cross-validation of the language model *Mammo-RoBERTa*. Left and right denote the models for predicting with the left and right breasts, respectively.

Implementation and hyper-parameters. We implement both the image model and the language model in PyTorch. All the mammograms are resized to 800×1024 pixels as a compromise between fast processing and high resolution. We initialize the network with the parameters pre-trained on the ImageNet classification task. We use several standard data augmentation techniques including random horizontal and vertical flipping, Gaussian noise, affine transformation (rotation, scaling) to avoid overfitting. We use Adam optimizer [14] to train Stage 1 for 50 epochs and set the initial learning rate to 0.0001, then reduce it to 1/5 of the initial value every 10 epochs. For the iterative training using the mixed supervised data (the combination of fully labeled and NLP labeled images), we perform alternate optimization for Q=2 iterations (no noteworthy improvement was observed after 2 iterations), with 15 epochs for each iteration. We set the learning rate for the classification branch 10 times to other layers of BTV. p_0 is initially set to 0.95 and decreases by 0.1 every 5 epochs to allow more NLP labeled examples to be fed into training. Through all the experiments, the batch size is 24 and the ratio of positive to negative samples is 1:5 in each mini-batch. We experiment with different combinations (α_t=$\{0.25, 0.5, 0.75\}$ and γ=$\{1, 2, 5\}$) and find that the default setting of α_t=0.25 and γ=2 in [20] is optimal. The weights of the loss components λ_1, λ_2, and λ_3 are empirically set to 0.3, 0.6, and 0.2, respectively.

Computation time. In terms of computation time, it takes about 15 hours to train the BTV for 50 epochs on 25% of the fully labeled data and 9 hours to train an additional iteration (15 epochs) on mix-supervised data, on 4 NVIDIA Quadro RTX 8000 GPUs. At the inference phase, an image goes through the BTV without the classification branch. On average, it takes about 0.1 seconds to predict the mass probability map and binary mask for one mammogram image.

4.3. Results of the language model

Based on the labeled mammograms we have on hand, we collected 3,753 mammography reports corresponding to these images. We randomly split them into three folds and evaluate the models with their corresponding hold-out set and get three results, each for one hold-out set. Then we aggregate these results as the evaluation result for the training

Labeled data	Weak data	FP @0.1	FP @0.2	FP @0.5	FP @1	Average Sen.
10%	✗	0.426	0.501	0.625	0.705	**0.564**
25%	✗	0.550	0.650	0.752	0.826	**0.695**
50%	✗	0.660	0.745	0.810	0.808	**0.756**
100%	✗	0.720	0.774	0.823	0.827	**0.786**
10%	✓	0.624	0.700	0.776	0.801	**0.725**
25%	✓	0.673	0.725	0.801	0.826	**0.753**
50%	✓	0.700	0.775	0.813	0.830	**0.780**
100%	✓	0.726	0.778	0.820	0.850	**0.794**

Table 3. The test set results of the BTV model with different proportions of fully labeled training data without (✗) or with (✓) weakly labeled data. Sensitivity at different false positives (FP) per image and average sensitivity are reported. The best result under each evaluation metric is underscored.

data. We evaluate the language model Mammo-RoBERTa with widely used F1 score and accuracy using a default threshold of 0.5 for quantitative measurement, though we adopted the probabilities from prediction as soft-labels instead of hard-labels after thresholding. We first calculate the F1 score of each label (mass or no mass) and then compute the average F1 as well as the weighted F1 as our metrics. Weighted F1 means the average F1 score is weighted by the number of samples of each label.

As can be seen from Table 2, our language classification model achieves very high accuracy (>0.98) in predicting the presence or absence of mass in text reports. The average F1 scores (>0.95) and weighted F1 scores (>0.98) also show that the performance of our language classification model is very robust in general. All these results demonstrate that the weakly labels extracted by our NLP model provide informative image-level supervision to the image model with minimum noise.

4.4. Evaluation

Evaluation metrics. We use the widely used Free-response Receiver Operating Characteristic (FROC) curve to evaluate the mass detection performance [2]. More specifically, we measure the sensitivities at 0.1, 0.2, 0.5, and 1 false positive per image (FPPI) to show the recall at different precision levels. There is a trade-off between a higher recall/sensitivity (towards 0% false negatives) and FP rate. Too many FPs might distract the radiologists. The average of these values is referred to as average sensitivity. Following prior work [2, 36], a mass is successfully detected if the intersection over the union (IoU) of a predicted mass region and the ground-truth mass mask is greater than 0.2.

Results using different proportions of fully and weakly labeled training data. We first evaluate the performance of the BTV model with different proportions of fully labeled training data without weakly labeled data on the test set. Detailed results are shown in row 2 to 5 of

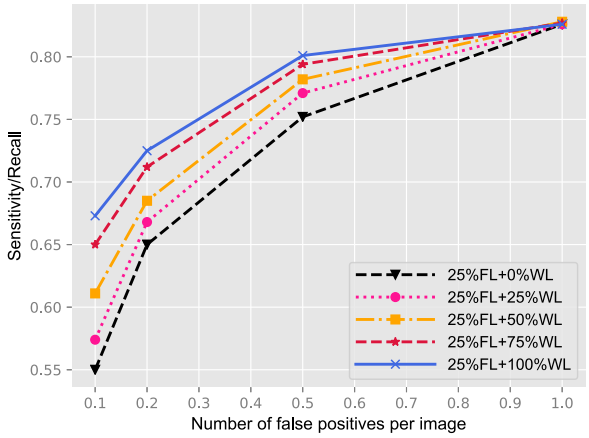

Figure 4. FROC curves of different proportions of weakly labeled training data when the amount of fully labeled data is fixed at 25%. FL: fully labeled by radiologists. WL: weakly labeled by NLP.

Table 3. As can be seen from these results, the performance improves with more fully labeled training images in general. When there are only a very limited number (*e.g.*, 10%) of fully labeled images, adding extra fully labeled data (*e.g.*, to 25%) markedly improves the performance. But when we have a mid-size (*e.g.*, 50%) labeled data, the benefit of adding extra annotations (*e.g.*, to 100%) is less obvious. These findings are consistent with previous work [18] in that fully labeled data is essential for detection/segmentation. We then evaluate the performance of the semi-supervised model by training with weakly labeled data using different proportions of fully labeled data. Row 6 to 9 of Table 3 demonstrate the results. Adding weakly labeled data using our iterative mining strategy consistently improves the performance from the baseline BTV model. Even though the baseline model trained with a small number (*e.g.*, 10%) of labeled images is less accurate, adding large-scale weakly labeled data immensely increases the performance. We can see that training using 25% (or 50%) of the fully labeled images together with the weakly labeled data, our semi-supervised model achieves performance on par with training using 50% (or 100%) of the fully labeled images. This suggests that the proposed semi-supervised learning model can effectively reduce the effort of manual annotations from medical experts. When trained with 100% of the fully labeled data, providing weakly labeled data still improves the model with a small margin. The curves in Figure 4 show the performance trends of introducing 0%, 25%, 50%, 75%, and 100% weakly labeled data when fully labeled training data is fixed at 25%. They demonstrate that the performance of our self-training method can be persistently increased with more weakly labeled data available.

Soft labels vs. hard labels. We investigate the impact of using soft labels in this study. We compare it with using hard labels by taking thresholds from soft labels. For the NLP labels, we set the threshold to 0.5 and binarize

mask	NLP	FP@0.1	0.2	0.5	1	Ave.
●	●	0.668	0.720	0.798	0.806	**0.748**
●	○	0.680	0.732	0.809	0.828	**0.762**
○	●	0.671	0.724	0.798	0.819	**0.753**
○	○	0.691	0.737	0.815	0.832	**0.769**

Table 4. Comparison of hard (●) versus soft labels (○) on the validation set using weakly labeled images with 25% labeled images.

Method	Data	FP@0.1	0.2	0.5	1	Ave.
Global [29]	WL only	0.475	0.530	0.551	0.605	**0.540**
Glb+Local [29]	WL only	0.512	0.563	0.624	0.687	**0.597**
Mixed [23]	All	0.608	0.690	0.741	0.794	**0.708**
MIL [18]	All	0.672	0.732	0.779	0.828	**0.753**
TCSM [17]	All	0.675	0.740	0.800	0.820	**0.759**
Ours	All	0.726	0.778	0.820	0.850	**0.794**

Table 5. Comparisons with state-of-the-art weakly- and semi-supervised learning methods in the medical imaging domain.

the labels into 0 and 1 (negative and positive mass image). For the pseudo masks, we take the binary masks after the Dense CRF layer instead of the probability maps as supervisions. The soft-label loss function is converted to the standard hard-label version for each component. When 25% fully labeled images are trained together with the weakly labeled data, the comparison of results using hard labels versus soft labels on the validation set is shown in Table 4. Results show the benefit of using soft labels and soft-label loss functions compared to their hard-label counterpart. Although some inaccurate NLP labels may mislead the image model, it might be partly alleviated by using the soft labels.

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

We present a comparison with state-of-the-art semi-supervised 2D medical image detection/segmentation approaches [18, 23, 17] using all our training images in Table 5, by reproducing the results on our test set. Li et al. [18] utilize a unified framework for disease localization and classification in a multiple instance learning (MIL) manner. A patch slicing layer is inserted before the recognition network to capture the local disease information from small patch grids. Mlynarski et al. [23] jointly train a framework for segmentation and classification to make full use of mixed supervised data (fully-labeled and weakly labeled). Li et al. [17] develop a self-ensembling approach that encourages a consistent prediction of the network for the same unlabeled input under various perturbations. A weighted combination of supervised and unsupervised losses is designed to optimize the semi-supervised segmentation framework. It is designed for the task where each image has a foreground. It does not use image-level labels and suffers from a bad local minimum. We also include the results from a weakly supervised localization model GMIC [29] which trained only with weak image-level labels for mass detection in mammograms. GMIC

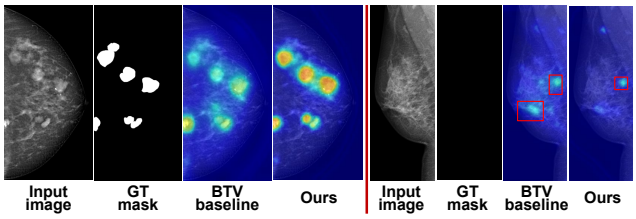

Figure 5. Two visualized exemplar mass detection results of our method. Detailed illustrations are described in Section 4.5.

contains a low-capacity global network operating at a whole mammogram and a high-capacity local network focusing on small patches. The final representation fusion is used to train a globally-aware multiple instance classifier. For all these methods, we replace their backbone DNNs with the revised HRNet as used in our model. As shown in Table 5, our method achieves the best performance compared with other methods. The comparisons with the weakly supervised model [29] in Table 5 and Table 3 show that fully labeled data is still essential in training a good model.

Two sets of qualitative results of our method are displayed in Figure 5. Each set contains the input mammogram, the ground-truth mass masks, the image overlay heatmaps (probability maps) by the supervised baseline, and the semi-supervised method, from left to right. In the first example, our iteratively trained semi-supervised model improves the baseline method with clearer boundaries and more confident predictions (the warmer color the larger score). It is clear in the second example when there is no mass in the image, the semi-supervised model generates less false positives (shown in red boxes) than the baseline model. This advantage owes to our iterative training using self-selected samples to suppress false positives.

5. Conclusion

In this paper, we studied a challenging problem in annotation-efficient medical imaging: how to leverage unlabeled or weakly labeled data. We developed an NLP model to extract probabilistic image-level labels from radiology reports as soft labels. Based on the fully and weakly labeled data, we designed a sample mining strategy for alternate training of a semi-supervised learning framework. The proposed method takes advantage of soft labels generated by the NLP and pixel-level prediction, and then models them as a means of uncertainties for weak data. It demonstrates better mass detection accuracy than the supervised baseline and other comparing methods. Future work includes extracting more information (*e.g.*, coarse locations of lesions) from report to guide the learning of the image model as a whole, and investigating self-supervised pre-training from unlabeled data to improve diagnostic accuracy.

Acknowledgements: We would like to acknowledge Tian Xia from PAII Inc. for helpful discussions.

References

[1] Self-transfer learning for weakly supervised lesion localization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

[2] Richa Agarwal, Oliver Diaz, Xavier Lladó, Moi Hoon Yap, and Robert Martí. Automatic mass detection in mammograms using deep convolutional neural networks. *Journal of Medical Imaging*, 6(3):031409, 2019.

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

[4] Ran Bakalo, Rami Ben-Ari, and Jacob Goldberger. Classification and detection in mammograms with weak supervision via dual branch deep neural net. In *IEEE International Symposium on Biomedical Imaging*, pages 1905–1909, 2019.

[5] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797, 2020.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

[7] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2020.

[8] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. Pre-training with whole word masking for chinese bert. *arXiv:1906.08101*, 2019.

[9] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[11] Wenchong He and Zhe Jiang. Semi-supervised learning with the em algorithm: A comparative study between unstructured and structured prediction. *IEEE Transactions on Knowledge and Data Engineering*, (01):1–1, aug 2020.

[12] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017.

[13] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations(ICLR)*, 2015.

[15] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 109–117, 2011.

[16] Haohan Li and Zhaozheng Yin. Attention, suggestion and annotation: A deep active learning framework for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–13, 2020.

[17] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, Lei Xing, and Pheng-Ann Heng. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–12, 2020.

[18] Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Li Fei-Fei. Thoracic disease identification and localization with limited supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[19] Gongbo Liang, Xiaoqin Wang, Yu Zhang, and Nathan Jacobs. Weakly-supervised self-training for breast cancer localization. In *Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1124–1127, 2020.

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[21] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[22] Yuhang Liu, Fandong Zhang, Qianyi Zhang, Siwen Wang, Yizhou Wang, and Yizhou Yu. Cross-view correspondence reasoning based on bipartite graph convolutional network for mammogram mass detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3812–3822, 2020.

[23] Pawel Mlynarski, Hervé Delingette, Antonio Criminisi, and Nicholas Ayache. Deep learning with mixed supervision for brain tumor segmentation. *Journal of Medical Imaging*, 6(3):034002, 2019.

[24] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1742–1750, 2015.

[25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 91–99, 2015.

[26] Dezső Ribli, Anna Horváth, Zsuzsa Unger, Péter Pollner, and István Csabai. Detecting and classifying lesions in mam-

mograms with deep learning. *Scientific Reports*, 8(1):1–7, 2018.

[27] 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*, pages 234–241, 2015.

[28] Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. Semi-supervised self-training of object detection models. In *IEEE Workshops on Applications of Computer Vision*, volume 1, pages 29–36, 2005.

[29] Yiqiu Shen, Nan Wu, Jason Phang, Jungkyu Park, Kangning Liu, Sudarshini Tyagi, Laura Heacock, S. Kim, Linda Moy, and Kyunghyun Cho. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *arXiv preprint arXiv:2002.07613*, 2020.

[30] Yuxing Tang, Josiah Wang, Boyang Gao, Emmanuel Dellandréa, Robert Gaizauskas, and Liming Chen. Large scale semi-supervised object detection using visual and semantic knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2119–2128, 2016.

[31] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[32] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2097–2106, 2017.

[33] Martin J Willemink, Wojciech A Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R Folio, Ronald M Summers, Daniel L Rubin, and Matthew P Lungren. Preparing medical imaging data for machine learning. *Radiology*, 295(1):4–15, 2020.

[34] Ziang Yan, Jian Liang, Weishen Pan, Jin Li, and Changshui Zhang. Weakly-and semi-supervised object detection with expectation-maximization algorithm. *arXiv preprint arXiv:1702.08740*, 2017.

[35] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 399–407, 2017.

[36] Zhicheng Yang, Zhenjie Cao, Yanbo Zhang, Mei Han, Jing Xiao, Lingyun Huang, Shibin Wu, Jie Ma, and Peng Chang. Momminet: Mammographic multi-view mass identification networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 200–210, 2020.