DK-Consistency: A Domain Knowledge Guided Consistency Regularization Method for Semi-supervised Breast Cancer Diagnosis

Xiaozheng Xie
School of Computer Science and Engineering (SCSE)
Beihang University
Beijing, China
xiexzheng@buaa.edu.cn

Jianwei Niu
SCSE, Beihang University
Beijing, China
Hangzhou Innovation Institute
School of Information Engineering
Zhengzhou University
Zhengzhou, China
niujianwei@buaa.edu.cn

Xuefeng Liu
SCSE
Beihang University
Beijing, China
liu_xuefeng@buaa.edu.cn

Qingfeng Li
Hangzhou Innovation Institute
Beihang University
Hangzhou, China
liqingfeng@buaa.edu.cn

Yong Wang
Chinese Academy of Medical Sciences
and Peking Union Medical College
Beijing, China
drwangyong77@163.com

Shaojie Tang
Jindal School of Management
The University of Texas at Dallas
TX, USA
shaojie.tang@utdallas.edu

2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-0126-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/BIBM52615.2021.9669494

***Abstract*—The performance of deep learning models generally relies on large and high-quality labeled datasets. However, in medical domain, as labeling process is much more laborious and time-consuming, most medical datasets are much smaller compared with natural image datasets. To mitigate this weakness, recent researches in medical image analysis adopt semi-supervised learning methods, especially consistency regularization methods to learn from a large amount of unlabeled medical data. However, as these semi-supervised learning methods are originally designed for tasks of natural images, specific properties of medical domain are not fully investigated and utilized. In this paper, we present DK-Consistency, a domain knowledge guided consistency regularization method for semi-supervised breast cancer diagnosis in ultrasound images. In DK-Consistency, domain knowledge of medical doctors is first incorporated into the generation process of perturbed samples for each unlabeled image. Then consistency regularization is adopted to force the model to make consistent predictions for unlabeled images and their perturbed samples. Extensive experiments demonstrate that, by injecting domain knowledge, DK-Consistency significantly improves the diagnostic performance of breast cancer and outperforms many state-of-the-art semi-supervised methods.**

***Index Terms*—domain knowledge, consistency regularization, semi-supervised breast cancer diagnosis**

I. Introduction

In recent years, with the wide adoption of deep learning techniques, various deep learning models promote booming development of computer-aided diagnosis (CAD) for different diseases [1], [2]. However, small-scale labeled medical datasets remain the primary bottleneck in the development of good deep learning based CAD models. Unlike natural image datasets, the labeling process of medical datasets is much more laborious, time-consuming and highly dependent on the professional knowledge of doctors. For example, each image is typically labeled by multiple doctors to reduce the intra-class differences.

To overcome the above challenge, it is a common practice to adopt semi-supervised learning methods [3]–[5]. In this way, CAD models can be built from a large amount of unlabeled medical data. One popular technique in semi-supervised learning is the consistency regularization strategy [6]. Consistency regularization devotes to make consistent predictions for images under different perturbations (e.g., Gaussian noise and randomized data transformation). Consistency regularization based methods (abbreviated as 'consistency-based methods' hereinafter) have been proven to be effective in many medical image analysis tasks, including skin lesion classification [7], nucleus classification [8] and breast cancer diagnosis [9].

Despite commendable results, current attempts to apply consistency-based methods to medical image analysis do not fully investigate and utilize the special properties of medical domain. In particular, the most special property of medical domain can be represented as the knowledge of medical doctors, including the way they browse images, the particular areas they focus on, and features they give special attention. This kind of knowledge, called as medical domain knowledge, has been incorporated in various ways into deep learning models for many medical image analysis tasks, by which better performance is achieved [10].

Considering the above success of applying medical domain knowledge with deep learning models, it is natural to believe that the knowledge can also help semi-supervised learning

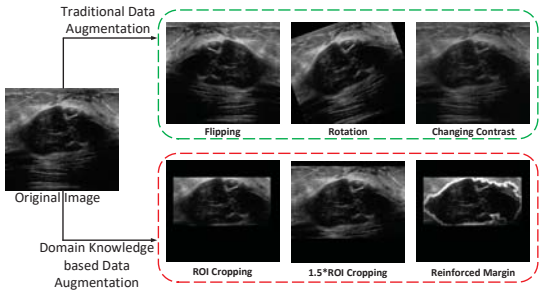

Fig. 1. The comparison of images generated using the traditional data augmentation method (upper branch) and the images generated by the augmentation method with domain knowledge (lower branch). Traditional data augmentation generally utilizes simple transformations (e.g., flipping, rotation, changing contrast). In contrast, samples generated by domain-knowledge based augmentation highlight areas where medical doctors pay more attention.

process to improve its performance. The way of incorporating medical domain knowledge lies in perturbed samples. In most consistency-based methods, perturbed samples of each image are generated via some data augmentation methods (e.g., rotation, flipping). These perturbed samples are used to help the model to learn important features of images. As features mentioned by medical domain knowledge can be more representative than random augmentation, the model using perturbed samples containing medical domain knowledge can have better performance.

Fig. 1 compares samples generated by traditional data augmentation method and by the method based on domain knowledge. Here the domain knowledge is represented as the tumor area, the tumor margin, and the surrounding contexture. We can find that, compared with traditional augmentation method, images generated with domain knowledge focus more on areas of tumor and tumor margin, which are believed to be more important for medical doctors to make decisions.

The above discussion leads to our proposed method called as DK-Consistency, a domain knowledge guided consistency regularization method for semi-supervised breast cancer diagnosis. Specifically, for each unlabeled medical image, three domain-knowledge augmented (DK-AUG) images are generated, each represents different features mentioned in medical domain knowledge. These DK-AUG images are combined with the image generated via RandAugment method [11] to obtain the final perturbed image denoted as the 'DK-guided image'. The model is then forced to make consistent predictions between each unlabeled image and its corresponding DK-guided image. In this way, the features that doctors focused and utilized can be learned. Meanwhile, supervised learning is also applied in DK-Consistency on labeled data to obtain reliable discriminative patterns.

We conduct extensive experiments on our breast ultrasound image dataset and a public dataset, and results show that DK-Consistency outperforms state-of-the-art methods in most evaluation metrics. The contributions are summarized as follows:

- We present a new domain knowledge guided data augmentation method to generate perturbed samples for semi-supervised breast cancer diagnosis.
- We propose a novel DK-Consistency paradigm by utilizing domain knowledge guided perturbed samples for the final consistent constraint.
- We conduct extensive experiments on two breast ultrasound image datasets, and experimental results demonstrate that DK-Consistency performs favorably against many state-of-the-art methods.

II. Related Work

We first introduce some consistency-based semi-supervised learning methods in medical area, and then recall researches that integrating medical domain knowledge into deep learning. Finally, we review recent work about breast cancer diagnosis.

A. Consistency-based Semi-supervised Learning for Medical Image Analysis

As a widely used semi-supervised learning technique, consistency-based methods enforce prediction consistency for inputs under different perturbations. Different variants of consistency-based approaches are proposed for natural image analysis tasks like the Π model [3], Mean Teacher [4], Virtual Adversarial Training (VAT) [12] and UDA [5]. These consistency-based methods have soon been applied to medical image analysis tasks. For example, by extending Π model, the transformation consistency strategy is utilized for the semi-supervised skin lesion segmentation [13]. Similarly, on the basis of Mean Teacher, a nuclei classification method enforces the local and global consistency of data to learn better features for classification task [8]. Moreover, adopting region of interest (ROI) consistency can also guide the network to focus on the brain region for brain MRI quality assessment [14]. Besides, the sample relation consistency paradigm is proposed to encourage the consistency of structured relation among different samples [7]. Meanwhile, increasing attention has been drown to different data augmentation methods in consistency training including RandAugment [11] and CTAugment [15].

Different from these methods, we introduce intrinsic features of medical images to generate perturbation samples for consistency training. These features are mainly represented as domain knowledge of medical doctors.

B. Integrating Domain Knowledge into Deep Learning for Medical Image Analysis

Nowdays, integrating domain knowledge into deep learning for medical image analysis attracts more attention. For example, a teacher-student curriculum learning strategy is proposed to mimic the from-easy-to-hard diagnostic pattern of doctors in breast screening classification [16]. In addition, the MommiNet is proposed to emulate how radiologists read mammogram images [17].

More recently, domain knowledge is also introduced into the semi-supervised learning process. For example, eight kinds of domain features are utilized to produce high quality images in the generative adversarial network (GAN) model for ultrasonography thyroid nodules classification [18]. Moreover, in

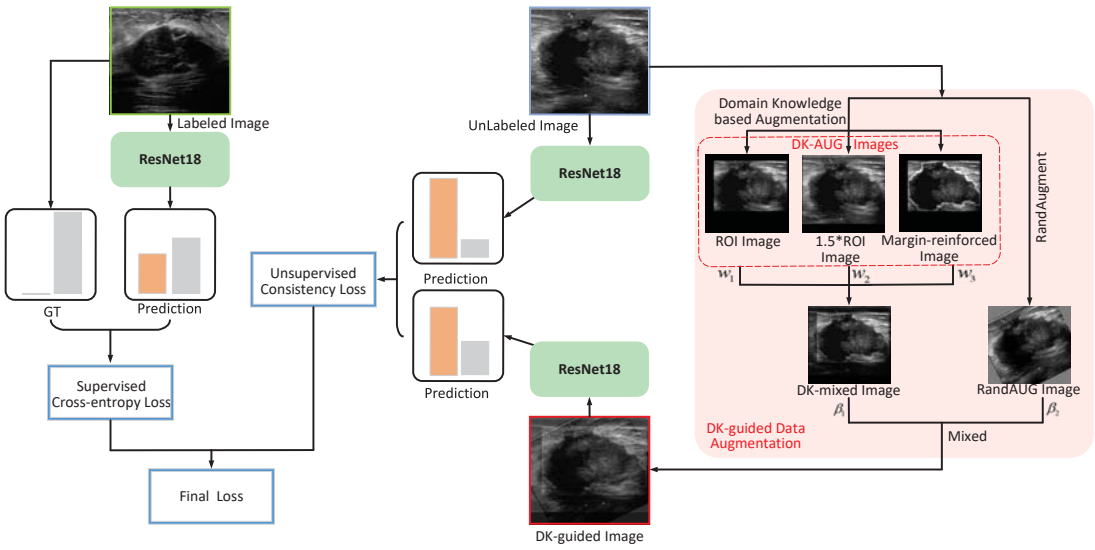

Fig. 2. Diagram of our DK-Consistency method. Both labeled and unlabeled data are used to train the same model (i.e., ResNet18) concurrently. In particular, cross entropy loss enforces the network to learn from labeled data between its ground truth label (GT) and the predicted result (Prediction), where the bar represents the probability of different categories. For the learning branch of unlabeled data, consistency prediction is constrained between the original image and its DK-guided image that generated using domain knowledge guided (DK-guided) data augmentation method.

[19], the semi-supervised domain adaptation approach (Dual-Teacher) is proposed for cardiac segmentation. However, the current researches are limited in either adopting GAN-based models or using some domain adaption methods.

In this paper, we introduce medical domain knowledge into the consistency regularization mechanism. Domain knowledge is directly used to generate more useful perturbed samples for unlabeled images, which simply enables the model to learn more informative features from unlabeled data.

C. Breast Cancer Diagnosis

As a most commonly occurring and high-fatal cancer, early detection and diagnosis of breast cancer is crucial to improve the survival rate of patients. Lots of efforts are dedicated to design CAD models [20], [21]. Early researches generally adopt the four-step diagnostic approach: pre-processing and segmentation, feature extraction, feature selection and classification [22]. More recent studies directly adopt deep learning models. For example, a Faster R-CNN based method is adopted to localize and classify masses from ultrasound images [20]. In addition, GoogLeNet is trained to differentiate benign and malignant tumors in [23]. Besides, some methods also incorporate different kinds of information, including margin information [24], handcrafted features [25], and information from other modalities [26] to improve the diagnostic performance.

Different from the aforementioned methods, in this paper, we focus on the semi-supervised breast cancer diagnosis. In addition, information of domain knowledge is introduced into perturbed samples and used to force model to learn more distinguishing features from unlabeled data.

III. Method

The overall architecture of our DK-Consistency is shown in Fig. 2. We can see that it consists of two branches: the supervised learning branch (the left part) and unsupervised learning branch (the right part). In particular, the former is to enforce the model to learn reliable discriminative patterns from labeled data, while the latter is utilized to encourage the model to make consistent prediction for unlabeled data and their perturbed samples (i.e., DK-guided images). The whole leaning process is regularized on the same model (i.e., ResNet18) with supervised and unsupervised loss concurrently.

A. Backbone of DK-Consistency Framework

In DK-Consistency, the training dataset mainly consists of the labeled set $D_L = \left\{\left(x_i, y_i\right)\right\}_{i=1}^{N_L}$ and the unlabeled set $D_{UL} = \left\{\left(x_i\right)\right\}_{i=1}^{N_{UL}}$. In particular, x_i represents the input image, y_i is the one hot ground-truth label (image category label in our method). N_L and N_{UL} are sizes of labeled set and unlabeled set, respectively. Normally, we assume that N_{UL} is much larger than N_L. In addition, $\hat{x}_i$ represents the DK-guided image of x_i, which is generated using DK-guided data augmentation method (will be described in Section III-B). During the training process, consistency-based learning framework explores the knowledge from all data concurrently. The overall objective function can be formulated as follows:

$$L = L_s + \lambda L_u, \tag{1}$$

where L_s and L_u represent the supervised loss and the unsupervised loss, respectively. The weighting factor λ is used to balance these two losses.

Specifically, L_s evaluates the network outputs on labeled data and is represented using cross-entropy:

$$L_s = \frac{1}{N_L} \sum_{x_i, y_i \epsilon D_L} H\left(y_i, f\left(x_i; \theta\right)\right), \tag{2}$$

where $H(p, q)$ is the cross-entropy between distributions p and q, $f(\cdot)$ denotes the classification network. θ is a set of network parameters. L_u evaluates the prediction consistency between unlabeled data and their DK-guided images:

$$L_u = \frac{1}{N_{UL}} \sum_{\substack{x_i \epsilon D_{UL}, \\ \hat{x}_i \sim q(\hat{x}_i | x_i)}} \left\| f(x_i, \theta') - f(\hat{x}_i, \theta) \right\|_2^2, \quad (3)$$

where θ' is a fixed copy of θ indicting that the gradient is not propagated through θ', as suggested by [12]. Additionally, $q(\hat{x}_i \mid x_i)$ is the DK-guided data augmentation transformation, in which domain knowledge is introduced into the generation process of $\hat{x}_i$.

B. DK-guided Data Augmentation

In this section, we introduce DK-guided data augmentation method, which is used to generate DK-guided images for unlabeled data. As shown in the red area of Fig. 2, the domain knowledge based augmentation method and the RandAugment method are applied to each unlabeled image, and their generated images are then combined to obtain the DK-guided image.

More specifically, the domain knowledge based augmentation method devotes to generate DK-AUG images containing domain knowledge of doctors. The knowledge is described as the consensus of the BI-RADS (Breast Imaging Reporting and Data System) [27], which is utilized by doctors to place abnormal findings into different categories. BI-RADS indicates some important sonographic characteristics including shape, margin and posterior features, to distinguish malignant tumors from benign ones. These features are generally closely related to some areas in images, such as areas of the tumor and its contexture, especially transition areas of the tumor margin.

We design three DK-AUG images to incorporate the above information. In particular, the ROI image contains ROI areas, the 1.5*ROI image contains both ROI and its surrounding areas, and the margin-reinforced image contains ROI image with reinforced pixels in margin areas.

The detailed process to generate the above DK-AUG images is shown in Fig. 3. We first adopt a fully automated segmentation method [28] to roughly locate the breast tumor in each image. Then the minimum bounding rectangle and 1.5 times bounding rectangle are calculated as the ROI area and 1.5*ROI area, respectively. Finally non-ROI area and non-1.5*ROI area cutout images are created as the ROI image and the 1.5*ROI image, respectively. For the margin-reinforced image, tumor contour is first extracted from the segmentation result, and dilated to obtain a ribbon mask with a certain width along the contour line. Then, based on the ROI image, pixels within margin areas are set to twice as before to obtain the margin-reinforced image. Subsequently, to better combine the information from these three DK-AUG images, we mix them by alpha compositing, and the element-wise convex coefficients in compositing are randomly sampled from Dirichlet distribution as in [29]. Lastly, the result is mixed again with the image generated via RandAugment method (denoted as the RandAUG image) to produce the final DK-guided image. Note that the RandAugment method can make the model more robust.

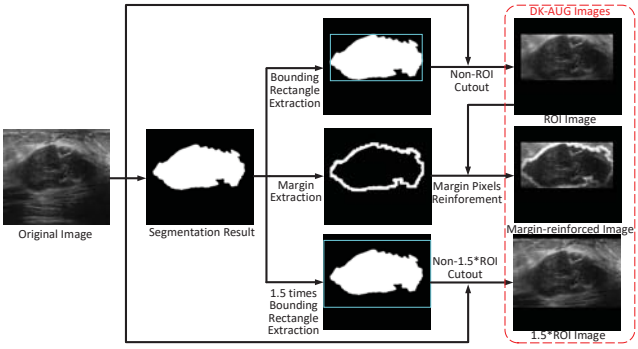

Fig. 3. The generation process of three DK-AUG images.

C. DK-Consistency Training Framework

During the training process, labeled images and unlabeled ones along with their DK-guided images are injected into the deep model (ResNet18) concurrently to extract features.

For labeled data, the prediction of model is compared with the ground truth label to enforce the correct classification. For the unlabeled data, the DK-Consistency paradigm encourages the same or similar prediction between the original image and its DK-guided images. In this way, category-related features can be learned from the branch of labeled data, while domain knowledge is exploited to address annotation scarcity and extract correct domain features from the side of unlabeled data. Thus, the whole learning process inherently focuses more on the features that domain knowledge indicated, by which the final performance can also be improved.

IV. Experiments

A. Datasets and Experimental Setup

Datasets Our proposed method is evaluated on two breast ultrasound image datasets, a private dataset and a public one. The essential information of these two datasets is as follows.

Our BUS2019 dataset is collected from the Cancer Hospital Chinese Academy of Medical Sciences from 2017 to 2019, and it mainly comprises 7318 images from 3946 patients. Within this dataset, 3596 images from 2039 patients are with benign tumors, while 3722 images from 1907 patients are with malignant ones. Samples in BUS2019 are collected from ultrasound systems with different vendors including PHILIPS, SIEMENS, HITACHI and so on. In particular, all images presented at least one tumor, and the benign or malignant label for each image is proved histopathology by biopsy.

The dataset is randomly divided into the training, validation and test sets according to the ratio 6:2:2. In detail, the training set consists of 4390 images (2158 benign images and 2232 malignant ones), while two groups of 1464 images (719 benign images and 745 malignant ones) are used as the validation set and the test set, respectively. Additionally, different percentages of images are also randomly selected from the training

TABLE I
QUANTITATIVE EVALUATION OF OUR METHOD UNDER DIFFERENT PERCENTAGES OF LABELED DATA

Method	Percentage		Metrics (%)				
	Labeled	Unlabeled	AUC	Accuracy	Sensitivity	Specificity	F1
Upper Bound	100%	0	87.46 ± 0.17	81.41 ± 0.30	83.06 ± 1.03	79.69 ± 1.29	81.97 ± 0.29
Baseline	5%	0	73.03 ± 0.85	67.57 ± 0.91	70.23 ± 2.94	64.81 ± 2.32	68.76 ± 1.37
DK-Consistency (ours)	5%	95%	75.72 ± 1.06	70.94 ± 0.65	74.20 ± 2.47	67.57 ± 1.92	72.20 ± 1.04
Baseline	10%	0	76.26 ± 1.56	69.46 ± 0.83	70.17 ± 1.87	68.71 ± 2.06	70.03 ± 0.93
DK-Consistency (ours)	10%	90%	80.23 ± 0.77	73.70 ± 0.71	75.33 ± 1.18	72.02 ± 1.74	74.46 ± 0.64
Baseline	20%	0	80.59 ± 1.00	74.74 ± 0.68	75.46 ± 1.69	73.99 ± 0.99	75.24 ± 0.88
DK-Consistency (ours)	20%	80%	85.20 ± 0.53	78.88 ± 0.45	80.49 ± 1.19	77.23 ± 1.66	79.51 ± 0.39
Baseline	30%	0	83.11 ± 0.50	75.92 ± 0.58	76.30 ± 0.68	75.52 ± 1.18	76.33 ± 0.50
DK-Consistency (ours)	30%	70%	86.25 ± 0.46	79.26 ± 0.32	81.02 ± 0.87	77.44 ± 0.89	79.90 ± 0.36

set as labeled ones, while the remaining images in training set are treated as unlabeled ones. It should be noted that images from the same patient are selected together into the same set or as labeled (or unlabeled) data in the training set.

The public available BUSI dataset [30] is collected from Baheya hostipal in 2018, which contains 780 images from 600 patients (437 benign, 210 malignant and 133 normal). The instruments used in the scanning process are LOGIQ E9 and LOGIQ E9 Agile ultrasound systems, and the image resolution is 1280×1024. As the domain knowledge focuses on benign and malignant images, normal images are not adopted. In addition, as the ratio of benign and malignant images in BUSI is 2:1, we adopt data augmentation method (i.e., rotating 10 degrees) on all malignant images to make a balanced dataset. Meanwhile, 5-fold cross-validation is used on this dataset.

Evaluation metrics Five evaluation metrics are adopted in all experiments, including AUC, accuracy, sensitivity, specificity and F1-score. It should be noted that each metric represents mean ± std over 5 trials.

Implementation details Our DK-Consistency method and other comparison methods are implemented with PyTorch library. To preserve the aspect ratio of tumor in the image, each image is first scaled to make the short edge to 224, and then center-cropped to 224*224. Additionally, the batch size of labeled images and unlabeled images are set to 32 and 960, respectively, the weighting factor of unsupervised loss λ is set to 1. We totally train 50 epochs for each classification task. The initial learning rate is 0.01875, and then decayed with a power of 0.1 after each 30 epochs. We use a SGD optimizer with the momentum hyperparameter set to 0.9. All experiments are trained on 2 GPUs of NVIDIA Tesla V100-PCIE-32GB.

B. Effectiveness of the DK-Consistency

To prove the effectiveness of DK-Consistency, we conduct experiments on different percentages (5%, 10%, 20% and 30%) of labeled images in our BUS2019 dataset. Quantitative results are shown in Table I. In particular, 'Upper Bound' refers to being trained with all 4390 labeled images in training set, while different baseline methods are trained only with the corresponding percentage of labeled data.

We can see that DK-Consistency achieves significant improvements over the corresponding baseline method. For example, in terms of AUC, DK-Consistency has 3.97% improvements compared with its baseline method using 10% of labeled data, and has 3.14% improvements when 30% of labeled data are utilized. In addition, DK-Consistency also shows lower standard deviations in most metrics (i.e., AUC, accuracy and F1 score). This indicates that better and more reliable performance can be achieved by introducing domain knowledge. Moreover, when using only 30% labeled data, DK-Consistency achieves the performance close to the upper bound method, in which the average of AUC and accuracy are 86.25% and 79.26%, respectively.

To give a more clear illustration, Fig. 4 compares the AUC (in the form of mean ± std) of the DK-Consistency with the baseline method at different percentages (the column of 'AUC' in Table I). We can clearly see the superior performance of the DK-Consistency over the baseline method at all percentages.

C. Comparison with Other Semi-supervised Methods

We compare the DK-Consistency with some state-of-the-art semi-supervised learning methods, including Mean Teacher [4], MixMatch [31], FixMatch [32], as well as the original UDA method [5]. For all of these methods, we adopt the same network structure and training protocol (i.e., data preprocessing, training epochs, optimizer, learning rate schedule). The comparison is conducted on 20% of labeled data (878 images), and quantitative results are listed in Table II. The best performance in each metric is shown in bold.

We can see from Table II that DK-Consistency substantially outperforms other methods on most metrics (i.e., AUC, accuracy and F1). In particular, DK-Consistency has the average AUC and accuracy of 85.20% and 78.88%, respectively, which

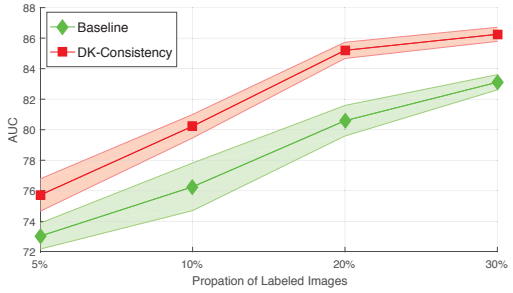

Fig. 4. Comparison the AUC under different percentages of labeled images.

TABLE II
COMPARISON WITH OTHER SEMI-SUPERVISED LEARNING METHODS

Method	Percentage		Metrics (%)				
	Labeled	Unlabeled	AUC	Accuracy	Sensitivity	Specificity	F1
Baseline	20%	0	80.59 ± 1.00	74.74 ± 0.68	75.46 ± 1.69	73.99 ± 0.99	75.24 ± 0.88
MeanTeacher [4]	20%	80%	83.33 ± 0.44	76.83 ± 0.63	79.28 ± 0.39	74.30 ± 0.98	77.69 ± 0.55
UDA [5]	20%	80%	83.30 ± 0.76	76.75 ± 1.12	79.65 ± 1.62	73.74 ± 1.45	77.71 ± 1.12
MixMatch [31]	20%	80%	84.46 ± 0.34	77.73 ± 0.55	**81.21 ± 1.35**	74.13 ± 1.71	78.77 ± 0.52
FixMatch [32]	20%	80%	84.44 ± 0.35	77.98 ± 0.35	76.08 ± 2.23	**79.94 ± 2.85**	77.85 ± 0.36
DK-Consistency (ours)	20%	80%	**85.20 ± 0.53**	**78.88 ± 0.45**	80.49 ± 1.19	77.23 ± 1.66	**79.51 ± 0.39**

are 0.76% and 0.9% improvements over FixMatch. We notice that although FixMatch achieves the highest specificity, it comes at the cost of a much lower sensitivity. The similar phenomenon can also be found in MixMatch (the highest sensitivity at the cost of a low specificity). Overall speaking, DK-Consistency boosts the performance in both sensitivity and specificity, which are 5.03% and 3.24% better than the baseline method. As an illustration, the AUC of DK-Consistency and other methods are shown in Fig. 5 as error bars. We can see that on average, DK-Consistency has better AUC values.

D. Ablation Studies

We also provide some ablation studies to further investigate the advantage of our DK-Consistency method. More specifically, we first analyze the effect of using DK-AUG images, then we compare the performance of different ways to introduce DK-AUG images into the perturbed sample generation process. The performance at these different conditions is summarized in Table III.

To prove the effectiveness of DK-AUG images, we compare three kinds of perturbed samples: (1) only using RandAUG images generated using method in [5] (indicated as 'RandAugment'), (2) only using DK-mixed images which are combined based on three DK-AUG images (indicated as 'dk3mix'), and (3) using DK-guided images in our DK-Consistency method.

The results of three methods, along with the baseline which only utilizes 20% labeled data, are shown in the upper parts of Table III. First, we can see that all these three augmentation methods have better performance than the baseline method. In addition, by using DK-AUG images, dk3mix can achieve comparative performance with the RandAugment method. More importantly, by combining DK-AUG images with samples generated by RandAugment, DK-Consistency achieves the top performance in major evaluation metrics (i.e., AUC and accuracy). This indicates that *domain knowledge does provide extra information that is not contained in samples generated by traditional random augmentation method*. This extra information helps the model to make better decisions.

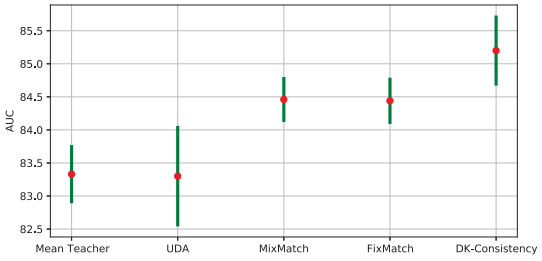

Fig. 5. AUC of different methods (in the form of error bar).

We also compare the performance of different ways in introducing DK-AUG images into the generation process of final perturbed samples. These methods are illustrated in Fig. 6 (a)-(e). Concretely, Fig. 6 (a) shows a method, denoted as 'dk3mix_aug', in which three DK-AUG images are first mixed and the resultant image is directly applied with RandAugment method. Fig. 6 (b) shows another method, denoted as 'dkaug3_mix', in which RandAugment method is first applied to three DK-AUG images and then the augmented results are mixed. In the method shown in Fig. 6 (c), denoted as 'dk3mix_img_aug', three DK-AUG images are first mixed, and then the resultant image is further mixed with the original input image, followed by the RandAugment method. In Fig. 6 (d) (denoted as 'dkaug3mix_imgaug'), the mixed image generated from Fig. 6 (b) is mixed with the image augmented from original image using RandAugment method. Fig. 6 (e) shows the DK-consistency method for better comparison.

Quantitative results of the methods show in Fig. 6 (a)∼(d) are listed in the lower part of Table III. It is observed that all these four methods improve the performance in some extent. In particular, 'dkaug3_mix' tends to improve sensitivity and specificity concurrently, but with large standard deviations. In addition, 'dk3mix_img_aug' highlights important areas by mixing DK-AUG images with the original image, which tends to improve sensitivity more than specificity. The performance of 'dkaug3mix_imgaug' is slightly lower than that in 'dkaug3mix'. The reason may be that the final perturbed sample used in the former condition (mixed image of 'dkaug3_mix' image and augmented original image) is a bit different with the original input image when compared with the latter condition.

Different from above conditions, by directly mixing the 'dk3mix' image with the augmented original input image (RandAUG image), our DK-Consistency method achieves the best performance among these conditions, which boosts the AUC and accuracy to 85.20% and 78.88%, respectively. Meanwhile, the method combines advantages of these two kinds of augmented images in a simple and elegant way, and it can also be directly applied to other similar scenarios.

E. Performance on Another Public Dataset

To evaluate the generalization of our DK-Consistency method, we also apply it on the public dataset BUSI [30].

TABLE III
ABLATION STUDY

Method	Percentage		Metrics (%)				
	Labeled	Unlabeled	AUC	Accuracy	Sensitivity	Specificity	F1
Baseline	20%	0	80.59 ± 1.00	74.74 ± 0.68	75.46 ± 1.69	73.99 ± 0.99	75.24 ± 0.88
RandAugment [11]	20%	80%	83.57 ± 0.93	76.95 ± 1.08	79.53 ± 1.69	74.27 ± 1.97	77.83 ± 1.04
dk3mix	20%	80%	83.09 ± 0.61	76.91 ± 0.43	**81.53 ± 0.75**	72.13 ± 1.09	78.23 ± 0.38
DK-Consistency (ours)	20%	80%	**85.20 ± 0.53**	**78.88 ± 0.45**	80.49 ± 1.19	77.23 ± 1.66	**79.51 ± 0.39**
dk3mix_aug	20%	80%	83.46 ± 0.59	77.55 ± 0.58	78.36 ± 1.72	76.72 ± 2.46	78.04 ± 0.47
dkaug3_mix	20%	80%	84.63 ± 0.75	78.37 ± 0.80	78.07 ± 2.53	**78.69 ± 2.19**	78.59 ± 1.05
dk3mix_img_aug	20%	80%	84.79 ± 0.30	78.18 ± 0.71	81.23 ± 1.41	75.02 ± 1.81	79.12 ± 0.66
dkaug3mix_imgaug	20%	80%	84.55 ± 0.54	78.14 ± 0.57	80.46 ± 1.12	75.74 ± 1.56	78.93 ± 0.51

Fig. 6. Different methods of incorporating DK-AUG images into the generation process of perturbed samples in ablation study.

The same pre-processing method and training protocol that used in our dataset are also adopted on BUSI. In addition, we also compare the performance of our DK-Consistency method with the method only using RandAUG samples (denoted as 'RandAugment'). Table IV shows the performance of these different methods under different percentages (10%, 15%, 20% and 30%) of labeled data settings on BUSI dataset. It should be noted that each value in the table represents the mean ± std result of three times of 5-fold cross-validation, and the best values in different labeled data settings are shown in bold. Additionally, 'Upper Bound' represents the method trained on all images from the training set of BUSI, and 'Baseline' methods under different percentages of labeled data settings are trained with the corresponding labeled data.

We can find that from Table IV, our DK-Consistency method shows better performance in most metrics when compared with the corresponding baseline method and the RandAugment method. For example, DK-Consistency boosts the average of accuracy to 78.83%, 79.75%, 80.23% and 81.10% under 10%, 15%, 20% and 30% labeled data settings, respectively. The performance is improved by 4.65%, 2.62%, 1.25% and 1.55% when compared with the corresponding baseline method, which also shows a trend of declining as the labeled data gradually increases. Meanwhile, DK-Consistency also achieves relatively smaller standard deviation when compared with the other two methods.

Besides, we can see that for both DK-Consistency and the RandAugment methods, the improvement in sensitivity is not as significant as that in specificity. Moreover, the RandAugment method has lower AUC values when compared with the corresponding baselines in 15%, 20% and 30% labeled data settings. We believe the reason can be attributed to the BUSI dataset itself. It is a small-scale and imbalanced dataset, with much fewer malignant images than benign ones. Implementing RandAugment method on such a dataset can introduce noises and make the model have difficulty in detecting malignant

TABLE IV
QUANTITATIVE EVALUATION ON THE PUBLIC DATASET BUSI

Method	Percentage		Metrics (%)				
	Labeled	Unlabeled	AUC	Accuracy	Sensitivity	Specificity	F1
Upper Bound	100%	0	92.29 ± 1.70	86.92 ± 1.90	79.37 ± 4.97	90.55 ± 2.43	79.72 ± 3.11
Baseline	10%	0	78.50 ± 2.34	74.18 ± 2.62	**66.98 ± 7.11**	77.61 ± 6.11	62.68 ± 2.73
RandAugment [11]	10%	90%	79.05 ± 4.75	76.70 ± 4.50	65.65 ± 7.13	82.01 ± 6.56	64.75 ± 5.81
DK-Consistency (ours)	10%	90%	**82.12 ± 2.86**	**78.83 ± 2.76**	66.19 ± 5.97	**84.89 ± 3.74**	**66.95 ± 4.54**
Baseline	15%	0	82.18 ± 3.00	77.13 ± 2.74	**69.84 ± 5.88**	80.63 ± 4.86	66.47 ± 3.27
RandAugment [11]	15%	85%	81.05 ± 3.68	79.04 ± 3.24	67.58 ± 4.99	84.56 ± 5.45	67.80 ± 3.83
DK-Consistency (ours)	15%	85%	**83.26 ± 2.32**	**79.75 ± 2.65**	67.46 ± 2.07	**85.95 ± 3.68**	**68.69 ± 2.95**
Baseline	20%	0	83.82 ± 4.18	78.98 ± 2.63	**70.00 ± 5.75**	83.30 ± 3.36	68.34 ± 3.87
RandAugment [11]	20%	80%	83.46 ± 3.41	79.64 ± 2.41	69.23 ± 6.57	84.65 ± 3.27	68.77 ± 4.05
DK-Consistency (ours)	20%	80%	**84.10 ± 3.67**	**80.23 ± 3.32**	69.05 ± 3.98	**85.61 ± 4.38**	**69.49 ± 4.33**
Baseline	30%	0	84.32 ± 2.83	79.55 ± 3.26	71.27 ± 4.80	83.53 ± 4.89	69.43 ± 4.06
RandAugment [11]	30%	70%	83.65 ± 2.92	79.76 ± 2.65	**71.65 ± 6.99**	83.66 ± 4.10	69.62 ± 3.84
DK-Consistency (ours)	30%	70%	**84.79 ± 2.63**	**81.10 ± 2.70**	69.52 ± 4.93	**86.65 ± 4.78**	**70.53 ± 3.34**

images, leading to a decrease in the sensitivity and AUC.

V. Conclusion

In this paper, we present a simple while effective semi-supervised method, DK-Consistency, to introduce medical domain knowledge into breast cancer diagnosis. Extensive experiments on our dataset and a public dataset demonstrate that, using domain knowledge to generate perturbed samples, DK-Consistency can help the model learn more distinguishing domain features from unlabeled data. We hope that DK-Consistency will encourage future research to incorporate domain knowledge into consistency-based semi-supervised method for different medical image analysis tasks.

Acknowledgment

This work was supported by National Natural Science Foundation of China (grant numbers 61976012, 61772060), Beijing Hope Run Special Fund of Cancer Foundation of China (grant number LC2019A01) and Beijing Dongcheng District Excellent Talent Training Subsidy Program.

References

[1] A. Esteva, B. Kuprel, R. A. Novoa, J. M. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[2] X. Han, J. Wang, W. Zhou, C. Chang, S. Ying, and J. Shi, "Deep doubly supervised transfer network for diagnosis of breast cancer with imbalanced ultrasound imaging modalities," in *Medical Image Computing and Computer Assisted Intervention*, 2020, pp. 141–149.

[3] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *International Conference on Learning Representation*, 2017.

[4] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in Neural Information Processing Systems*, 2017, pp. 1195–1204.

[5] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 6256–6268.

[6] P. Bachman, O. Alsharif, and D. Precup, "Learning with pseudo-ensembles," *Advances in neural information processing systems*, vol. 27, pp. 3365–3373, 2014.

[7] Q. Liu, L. Yu, L. Luo, Q. Dou, and P. Heng, "Semi-supervised medical image classification with relation-driven self-ensembling model," *IEEE Trans. Med. Imag.*, vol. 39, no. 11, pp. 3429–3440, 2020.

[8] H. Su, X. Shi, J. Cai, and L. Yang, "Local and global consistency regularized mean teacher for semi-supervised nuclei classification," in *Medical Image Computing and Computer Assisted Intervention*, 2019, pp. 559–567.

[9] X. Wang, H. Chen, H. Xiang, H. Lin, X. Lin, and P.-A. Heng, "Deep virtual adversarial self-training with consistency regularization for semi-supervised medical image classification," *Medical Image Analysis*, vol. 70, p. 102010, 2021.

[10] X. Xie, J. Niu, X. Liu, Z. Chen, S. Tang, and S. Yu, "A survey on incorporating domain knowledge into deep learning for medical image analysis," *Medical Image Analysis*, p. 101985, 2021.

[11] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical data augmentation with no separate search," *arXiv preprint arXiv:1909.13719*, 2019.

[12] T. Miyato, S. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1979–1993, 2019.

[13] X. Li, L. Yu, H. Chen, C.-W. Fu, and P.-A. Heng, "Semi-supervised skin lesion segmentation vis transformation consistent self-ensembling model," *arXiv preprint arXiv:1808.03887*, 2018.

[14] J. Xu, S. Lala, B. Gagoski, E. Abaci Turk, P. E. Grant, P. Golland, and E. Adalsteinsson, "Semi-supervised learning for fetal brain mri quality assessment with roi consistency," in *Medical Image Computing and Computer Assisted Intervention*, 2020, pp. 386–395.

[15] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel, "Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring," in *International Conference on Learning Representations*, 2020.

[16] G. Maicas, A. P. Bradley, J. C. Nascimento, I. Reid, and G. Carneiro, "Training medical image analysis systems like radiologists," in *Medical Image Computing and Computer-Assisted Intervention*, 2018, pp. 546–554.

[17] Z. Yang, Z. Cao, Y. Zhang, M. Han, J. Xiao, L. Huang, S. Wu, J. Ma, and P. Chang, "Momminet: Mammographic multi-view mass identification networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2020, pp. 200–210.

[18] W. Yang, J. Zhao, Y. Qiang, X. Yang, Y. Dong, Q. Du, G. Shi, and M. B. Zia, "Dscgans: Integrate domain knowledge in training dual-path semi-supervised conditional generative adversarial networks and s3vm for ultrasonography thyroid nodules classification," in *Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 558–566.

[19] K. Li, S. Wang, L. Yu, and P.-A. Heng, "Dual-teacher: Integrating intra-domain and inter-domain teachers for annotation-efficient cardiac segmentation," in *Medical Image Computing and Computer Assisted Intervention*, 2020, pp. 418–427.

[20] S. Y. Shin, S. Lee, I. D. Yun, S. M. Kim, and K. M. Lee, "Joint weakly and semi-supervised deep learning for localization and classification of masses in breast ultrasound images," *IEEE Trans. Med. Imag.*, vol. 38, no. 3, pp. 762–774, 2019.

[21] X. Tang, L. Zhang, W. Zhang, X. Huang, V. Iosifidis, Z. Liu, M. Zhang, E. Messina, and J. Zhang, "Using machine learning to automate mammogram images analysis," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2020, pp. 757–764.

[22] M. A. Mohammed, B. Al-Khateeb, A. N. Rashid, D. A. Ibrahim, M. K. Abd Ghani, and S. A. Mostafa, "Neural network and multi-fractal dimension features for breast cancer classification from ultrasound images," *Computers and Electrical Engineering*, vol. 70, pp. 871–882, 2018.

[23] S. Han, H.-K. Kang, J.-Y. Jeong, M.-H. Park, W. Kim, W.-C. Bang, and Y.-K. Seong, "A deep learning framework for supporting the classification of breast lesions in ultrasound images," *Physics in Medicine & Biology*, vol. 62, no. 19, pp. 7714–7728, sep 2017.

[24] X. Xie, J. Niu, X. Liu, Q. Li, Y. Wang, J. Han, and S. Tang, "Dg-cnn: Introducing margin information into cnn for breast cancer diagnosis in ultrasound images," *Journal of Computer Science and Technology*, 2020.

[25] H. Feng, J. Cao, H. Wang, Y. Xie, D. Yang, J. Feng, and B. Chen, "A knowledge-driven feature learning and integration method for breast cancer diagnosis on multi-sequence mri," *Magnetic Resonance Imaging*, vol. 69, pp. 40–48, 2020.

[26] R. K. Samala, H.-P. Chan, L. Hadjiiski, M. A. Helvie, C. D. Richter, and K. H. Cha, "Breast cancer diagnosis in digital breast tomosynthesis: effects of training sample size on multi-stage transfer learning using deep neural nets," *IEEE Trans. Med. Imag.*, vol. 38, no. 3, pp. 686–696, 2018.

[27] E. Mendelson, M. Bohm-Velez, W. Berg, G. Whitman, I. Feldman, and H. Madjar, *ACR BI-RADS Ultrasound, ACR BI-RADS® atlas, breast imaging reporting and data system*. Reston, VA: American College of Radiology, 2013.

[28] W. Gmez-Flores and B. A. Ruiz-Ortega, "New fully automated method for segmentation of breast lesions on ultrasound based on texture analysis," *Ultrasound in Medicine & Biology*, vol. 42, no. 7, pp. 1637–1650, 2016.

[29] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple data processing method to improve robustness and uncertainty," *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[30] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in Brief*, vol. 28, p. 104863, 2020.

[31] D. Berthelot, nicholas Carlini, I. Goodfellow, A. Oliver, N. Papernot, and C. Raffel, "Mixmatch: a holistic approach to semi-supervised learning," 2019, pp. 5050–5060.

[32] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 596–608.