Calibrate the Inter-Observer Segmentation Uncertainty via Diagnosis-First Principle

Junde Wu\textsuperscript{\circ}, Graduate Student Member, IEEE, Yu Zhang\textsuperscript{\circ}, Huihui Fang\textsuperscript{\circ}, Lixin Duan\textsuperscript{\circ}, Mingkui Tan\textsuperscript{\circ}, Member, IEEE, Weihua Yang\textsuperscript{\circ}, Chunhui Wang, Huiying Liu, Yueming Jin\textsuperscript{\circ}, Member, IEEE, and Yanwu Xu\textsuperscript{\circ}, Senior Member, IEEE

Abstract—Many of the tissues/lesions in the medical images may be ambiguous. Therefore, medical segmentation is typically annotated by a group of clinical experts to mitigate personal bias. A common solution to fuse different annotations is the majority vote, e.g., taking the average of multiple labels. However, such a strategy ignores the difference between the grader expertness. Inspired by the observation that medical image segmentation is usually used to assist the disease diagnosis in clinical practice, we propose the diagnosis-first principle, which is to take disease diagnosis as the criterion to calibrate the inter-observer segmentation uncertainty. Following this idea, a framework named Diagnosis-First Segmentation Framework (DiFF) is proposed. Specifically, DiFF will first learn to fuse the multi-rater segmentation labels to a single ground-truth which can maximize the disease diagnosis performance. We dubbed the fused ground-truth as Diagnosis-First Ground-truth (DF-GT). Then, the Take and Give Model (T&G Model) to segment DF-GT from the raw image is proposed. With the T&G Model, DiFF can learn the segmentation with the calibrated uncertainty that facilitates the disease diagnosis. We verify the effectiveness of DiFF on three different medical segmentation tasks: optic-disc/optic-cup (OD/OC) segmentation on fundus images, thyroid nodule segmentation on ultrasound images, and skin lesion segmentation on dermoscopic images. Experimental results show that the proposed DiFF can effectively calibrate the segmentation uncertainty, and thus significantly facilitate the corresponding disease diagnosis, which outperforms previous state-of-the-art multi-rater learning methods.

Index Terms—Diagnosis, medical image segmentation, multi-rater, uncertainty estimation, inter-observer variability.

I. INTRODUCTION

In most nature image segmentation tasks, the ground-truth of segmentation is unique and confident. However, in medical images, the target tissues/lesions can be very ambiguous, so the labels collected are commonly subjective. Therefore, collecting the annotations from several different clinical experts to mitigate individual bias is clinically essential. However, this comes at the cost of introducing inter-observer uncertainty to the annotations. Prior works [1], [2], [3] called it a ‘multi-rater problem’, which means that each instance of the dataset is annotated by several different raters. One multi-rater example of optic cup (OC) segmentation is shown in Figure 1. We can see a significant variance in the annotations provided by different clinical experts. With such a significant uncertainty on the segmentation label, it is hard to develop deep learning models as automated segmentation solutions and can not quantitatively analyze and evaluate the models.

When facing a multi-rater problem, a simple and popular approach is the majority vote. Although this fusion strategy is simple and easy to implement, it comes at the cost of ignoring the different expertness of multiple graders [4]. Recently, many works have been proposed to model this inter-observer variability [1], [5], [6], [7] from the multi-rater labels. It is
shown that a better modeling of the rater annotation features will improve the final performance. However, those models require the user-given rater expertise to obtain a unique segmentation. Since rater expertness is commonly unavailable in the practice, in most cases, they are still limited to predict the traditional majority vote.

Another branch of study proposed to jointly estimate the multi-rater expertness and the classification (or segmentation) to handle classification (or segmentation) uncertainty in their tasks \cite{8, 9, 10, 11, 12}. These methods are generally based on the Expectation-maximization (EM) algorithm. The main idea is to use the prediction in this training round to calibrate the ground-truth for the next-round supervision. In specific, most of those methods first predict an initial result (maybe classification or segmentation, based on their tasks) using the neural network, then assign the weights of raters based on this result, and then use the new weights to combine a new ground-truth to retrain the network. They will iterative this process many times until the convergence. In this process, the errors in their prediction will accumulate, lead to significant deviant in the end. These two branches of study show that we lack gold criteria to evaluate the expertness of each rater so as to calibrate the uncertainty accurately.

Clinically, the disease diagnosis is usually conducted based on critical biomarkers derived from an analysis of the images. For example, on fundus images, the vertical Cup-to-Disc Ratio (vCDR) parameter computed from the optic cup/disc (OD/OC) masks is one of the most important clinical parameters for the glaucoma diagnosis \cite{13}. In melanoma diagnosis, an asymmetrical and irregular shape of the skin lesions is a major biomarker indicating melanoma \cite{14}. This inspires us to take disease diagnosis as a criterion to estimate the rater expertness and fuse the multi-rater labels. The fused label then can be used as the ground-truth for the segmentation training.

Specifically, we propose Diagnosis-First Segmentation Framework (DiFF) to predict the calibrated segmentation masks based on the diagnosis-first principle. DiFF is implemented in two steps. The first step is to find the best fusion of the multi-rater labels for a diagnosis network, as what we did in our conference paper \cite{15}. In particular, we optimize the expertness maps for each rater so as the fusion of multiple labels could maximize the diagnosis performance of the diagnosis network. The label fused by these multi-rater expertness maps is named Diagnosis-First Ground-truth (DF-GT). To improve the generalization of DF-GT toward different diagnosis networks, we further propose Expertness Generator (ExpG) to generate the expertness maps in the optimization process.

In the second step, we propose Take-and-Give Model (T&G Model) to segment DF-GT from the raw images. T&G Model is designed to integrate the diagnosis knowledge into the segmentation network, so as to better capture the segmentation features related to the diagnosis. The experiment shows that by adopting proposed DiFF, the estimated segmentation masks can significantly improve the performance of diagnosis, which outperforms SOTA multi-rater learning methods.

In brief, the contributions of this paper can be summarized as follow:

- To our knowledge, we are the first to address the multi-rater problem by using disease diagnosis as the criterion. According to the diagnosis-first principle, we propose a novel deep learning framework called DiFF to calibrate the inter-observer uncertainty to facilitate the disease diagnosis.
- As a part of DiFF, we propose a novel strategy to fuse the multi-rater labels through diagnosis features. The fused ground-truth is called DF-GT. In this process, we further propose ExpG to eliminate the high-frequency components to improve the generalization.
- We propose T&G Model in DiFF to integrate the diagnosis knowledge into the segmentation network. T&G Model can select and reinforce the diagnosis-related features by attentive feature interaction.
- We verify the proposed models on three different medical segmentation tasks with different image modalities. The experiment shows that the segmentation results gain high diagnosis performance and consistently outperform previous SOTA multi-rater learning methods.

\section{Method}

\subsection{A. Motivation}

Clinically, the segmentation of lesions/tissues can significantly facilitate the disease diagnosis. The same thing can also be observed in deep learning based automated diagnosis models. Many prior works have shown that segmentation information can bring diagnosis networks with solid improvement \cite{16, 17, 18, 19, 20, 21}. The common practices include the input concatenation, region of interest (ROI) extraction \cite{16, 20}, channel attention \cite{17, 19}, vision transformer based feature fusion \cite{21}, and transfer learning \cite{18}. However, when the segmentation contains underlying inter-observer uncertainty, the relationship between them comes to be more complicated.

Raters tend to be systematically biased towards certain cases. The results given by raters depend on the images and their own experience and expertise. Due to differences in their experience and expertise, they often produce results that are more consistent with their past experience and expertise. For example, if we take the average of the results given by all raters as a standard, a less experienced expert tends to be more conservative, resulting in a larger annotated lesion area. While some expert may provide annotations with an area much smaller than that of others. We conducted a statistical analysis of the OD/OC annotations provided by different experts. With
the average area of all OC annotations set as 1, the OC annotations of Rater 1 (R1) through Rater 6 (R6) had areas of 1.05, 1.06, 0.90, 1.33, 0.83, and 0.83. As shown in the Fig. 2, the annotations of OC by Rater 4 (R4) are relatively larger, whereas those by Rater 5 (R5) and Rater 6 (R6) are generally smaller.

Therefore, the segmentation of each rater may have different effects on the disease diagnosis. Thus, fusing the multi-rater segmentation in different ways will also differently affect the diagnosis performance. In order to quantitatively analyze these effects on the diagnosis, we perform a preliminary experiment with an OD/OC segmentation setting on the REFUGE-2 benchmark [22].

We train aforementioned segmentation-assisted diagnosis networks, which are denoted as DL-CAT [23], DL-ROI [20], DL-ATT [19], DL-SSL [18], and DL-SeAT [21] to diagnose glaucoma from the fundus images, with optic OD/OC segmentation masks as the assisted information. The network is trained to predict the diagnosis labels, of which 0 refers to nonglaucoma and 1 refers to glaucoma. We then provide the models with different segmentation masks annotated by seven different raters, respectively. We use ResNet50 as the diagnosis backbone in all these methods for a fair comparison. The final diagnosis performance measured by AUC score (%) is shown in Figure 3.

Figure 3 shows that different raters’ segmentation will have different effects on the diagnosis. It is clear that Rater 2 (R2) and Rater 4 (R4) improve diagnosis performance more than the other raters, and R7 brings the least improvement compared with the others. Such a general pattern can be consistently observed on all kinds of segmentation-assisted diagnosis methods, except DL-ROI, which takes no significant improvement compared with NoMask. These results demonstrate that some raters segmented the images with particular patterns, which can be more beneficial to the diagnosis.

These specific preferences can also be modeled by the neural networks, which is in line with the previous findings [1]. Based on these observations, it can be inferred that training a segmentation network that can recognize and enhance the diagnosis-first rater preferences will help to calibrate the uncertainty in a way conducive to the disease diagnosis.

It can be proven that incorporating diagnosis result aids in reducing uncertainty in the medical segmentation task. The uncertainty of the task can be represented by the information entropy $H(A)$, where $A$ represents the known information as well as the information to be solved, such as the raw images and annotations from different raters. Generally, a lower information entropy corresponds to lower task uncertainty. Introducing more information to construct conditional entropy is a common approach to reducing uncertainty. With the diagnosis result $y$, the information entropy of this task can be expressed as $H(A|y)$. The relationship between $H(A|y)$ and $H(A)$ can be expressed as follows:

$$H(A|y) = H(A) - I(A, y)$$ (1)

where $I$ represents the mutual information, which is greater than or equal to 0. For this task, the mutual information can only be 0 if the diagnostic result $y$ is completely independent of the information $A$. According to the above analysis, it is not possible for mutual information value to be 0. Therefore, when we introduce the diagnostic result $y$ into this task, the entropy decreases, indicating a reduction in uncertainty.

**B. Diagnosis-First Segmentation Framework**

Inspired by the previous experiment and analysis, we intend to find a specific fusion of the multi-rater segmentation labels, which is able to maximize the diagnosis performance. Then, by training the segmentation network on such a kind of ground-truth, we can learn to predict the diagnosis-first segmentation masks from the raw images.

Following this idea, we propose DiFF, and its overall pipeline is shown in Figure 4 (a). Firstly, a weighted fusion of multi-rater segmentation labels, called DF-GT is learned from a diagnosis network, as shown in Figure 4 (b). Then, DF-GT can be used as the ground-truth to train the segmentation network.

During the training, the raw image $x$ is first sent to the segmentation encoder and diagnosis network to get the segmentation features and diagnosis features. Then, the T&G Model connects the features in the segmentation encoder, the diagnosis network, and the segmentation decoder to instill the prior diagnosis knowledge into the segmentation network. The binary cross entropy (BCE) loss is adopted as the loss function:

$$L = L_{bce}(DF-GT, M).$$ (2)

where $M$ represents the estimated mask of the segmentation decoder, which is also the final prediction of the method.

In following, we will introduce DF-GT learning process and T&G Model in detail.
Learning DF-GT : To fuse multiple labels to one ground-truth, we weight the sum of the multiple labels by the expertness of each rater:

\[
\text{Groundtruth} = s \odot m = \sum_{i=1}^{n} s_i \ast m_i,
\]

where \( \ast \) denotes the element-wise multiplication, \( s_i \) and \( m_i \) are the annotation and the expertness map of the rater \( i \), respectively. Each pixel value in \( s_i \) is either 0 or 1, where 1 represents that the pixel corresponds to a specific organ or tissue.

\( m \) is treated as a pixel-wise weight, which can represent the confidence level of the segmentation annotations produced by \( n \) raters on a pixel level, therefore, we name it as the expertness maps. Then, the Groundtruth can be obtained through the weighted average of different rater annotations. For one specific pixel, the sum of the weight values for different annotations is equal to 1. To achieve that, we apply the Softmax operation on \( m \) to normalize the weights in practical applications, which also enables the Groundtruth to be interpreted as a probability value between [0,1].

The rater expertness in this paper is obtained through the disease diagnosis performance. In other words, the rater who contributes more to the correct diagnosis would be given higher expertness. Toward that end, we take the multi-rater expertness maps as the learnable variables to maximize the diagnosis performance of a segmentation-assisted diagnosis network. The network can be implemented in any way. For convenience, we use basic DL-CAT as an example, i.e., a standard classification network inputted by a concatenation of the raw image and the segmentation mask. We pre-train the network on the training set of the dataset.
The details of learning DF-GT are shown in Figure 4 (b). Formally, consider each raw image \( x \in \mathbb{R}^{h \times w \times c} \) is annotated by \( n \) raters, resulting in \( n \) segmentation masks \( s \in \mathbb{R}^{h \times w \times h} \), and the image is diagnosed as \( y \in [0, 1] \). Let \( \theta \) denotes the set of diagnosis model parameters. \( L(\theta, x, y) \) denotes the loss function of a standard classification task. Our goal is to find an optimal expertness map \( m \in \mathbb{R}^{h \times w \times c} \) by solving the following MAP (Maximum A Posteriori) problem:

\[
\begin{align*}
m^* &= \arg \max_{m} \mathbb{P}(m|\theta, x, s, y) \\
&= \arg \max_{m} \frac{\mathbb{P}(y|\theta, x, s, m)\mathbb{P}(\theta, x, s, m)}{\mathbb{P}(y, x, s)} \\
&= \arg \max_{m} \frac{\mathbb{P}(y|\theta, x, s, m)\mathbb{P}(\theta, x, s, m)}{\mathbb{P}(\theta, x, s)} \\
&= \arg \min_{m} -\log(\mathbb{P}(y|\theta, x, s, m)) - \log(\mathbb{P}(m)) \tag{4}
\end{align*}
\]

where \( \mathbb{P} \) represents probability. The second term in Eqn.(4) represents the prior probability of \( m \). In most MAP estimations, it is not straightforward to specify the prior probability distribution. Therefore, it is common to introduce some form of regularization to constrain the parameters involved in the prior, in order to avoid certain issues.

First, we can consider only the first term of Eqn. (4). In this case, the problem can be represented as follows:

\[
\begin{align*}
m^* &= \arg \min_{m} L(\theta, x \oplus [s \circ m], y), \tag{5}
\end{align*}
\]

where \( \oplus \) denotes concatenation operation, \( m^* \) denotes the optimal expertness maps. As mentioned above, integrating the segmentation results with the raw image for disease diagnosis has been extensively explored. However, this is not the main focus of this study. Therefore, in Eqn. (5), we chose to use the simplest concatenation method to implement the fusion.

According to Eqn. (5), we are actually finding expertise maps that can minimize the diagnosis loss. We name the fused ground-truth under these expertise maps, i.e., \( s \circ m \) as DF-GT. We adopt gradient descent to solve Eqn. (5):

\[
\begin{align*}
m^{t+1} = m^t + \alpha \nabla_m L(\theta, x \oplus [s \circ m], y), \tag{6}
\end{align*}
\]

where \( \alpha \) is the learning rate.

The optimized results in Eqn. (6) can improve the diagnosis performance to a high level. But since it is optimized toward one specific diagnosis network, it will not be general enough. At the same time, due to the lack of constraints on \( m \) imposed by the second term of Eqn. (4), we cannot ensure that \( m \) will converge to a reasonable solution. The visualization results also show it suffers heavily from high-frequency noises (An example of OD/OC segmentation is shown in Figure 7). The latest findings suggest that these high-frequency components have a close relationship with the generalization capability [24, 25].

Therefore, a possible approach to make \( m \) converge to a reasonable solution and improve the generalization of DF-GT is to constrain the high-frequency components in the optimization process, which corresponds to the second term of the Eqn. (4).

On the other hand, the high-frequency components observed in our samples differ from random patterns like salt-and-pepper noise. These are, in fact, a consequence of the convolutional network structure. Specifically, these artifacts result from backpropagating through the convolution layer directly to the input in our method. Previous research, such as [26], has indicated that deconvolution or transposed convolution can create overlapping artifacts. When computing gradients of convolution in backpropagation, these transposed-convolution artifacts are accumulated, which ultimately manifest as high-frequency components on the input. This is the key reason for the presence of high-frequency information when generating expertise maps through backpropagation. Traditional denoising algorithms are struggle to address these specific high-frequency components.

We tried several methods to constrain the high-frequency components, including Transformation Robustness (TransRob), Fourier Transform (Fourier), and the proposed Expertness Generator (ExpG). Among them, Transformation Robustness constrains high-frequency gradients by applying small transformations to the expertness map before optimization. In practice, we rotate, scale and jitter the maps. Fourier Transform transforms the expertness map parameters to the frequency domain, thus decorrelated the relationship between the neighborhood pixels. Unlike the aforementioned non-learnable methods, the proposed ExpG is implemented by a tiny CNN based pixel generator, which consists of four CNN layers.

The four-lyaer ExpG can use the CNN’s inductive bias to constrain the high-frequency components. Previous studies [27], [28], [29] have demonstrated that neural networks are more inclined to learn and generate low-frequency information. Therefore, feeding a CNN continuous coordinates can constrain it to produce a continuous, low-frequency map, which underpins the rationale behind our ExpG design. By understanding how high-frequency components are introduced and amplified through the backpropagation of a fixed network, and considering the inductive bias of CNNs, we implemented ExpG not as an ad-hoc solution, but as a strategic enhancement to elevate our model’s overall performance. The similar technique, such as using implicit neural networks to represent variables, has been extensively studied and explored, including mapping two-dimensional coordinates [30] or three-dimensional coordinates [29] to RGB values.

An illustration of ExpG is shown in Figure 4 (b). The network scans one pixel at a time. For each pixel, it predicts the pixel value with the position of the pixel. The input of ExpG is the coordinate vector \((i, j)\), and the output is the pixel value. ExpG is optimized to generate expertise maps that can minimize the diagnosis loss. Denote the parameters of ExpG as \( \phi \), our goal is to solve:

\[
\begin{align*}
\arg \min_{\phi} L(\theta, x \oplus [s \circ \text{ExpG}_\phi(i, j)]_{i=1}^{1-w}, y), \tag{7}
\end{align*}
\]

where \( \{\text{ExpG}_\phi(i, j)\}_{i=1}^{1-w} \) denotes generated expertness map with size \( h \times w \times h \) by ExpG. Since the continuity of the neural network mapping function, similar inputs tend to cause similar outputs, which leads the element values in expertise maps to
variant smoothly between the positions and thus eliminate the high-frequency components.

2) T & G Model: In order to predict diagnosis-first ground-truth (DF-GT) from the given raw images for inference, we need to train a segmentation network under its supervision. However, the standard segmentation network will show subpar performance on DF-GT. That is because first, the standard segmentation network can not access diagnosis information. Without this prior knowledge, the segmentation network is hard to capture the features related to the diagnosis, thus fails to learn DF-GT. Meanwhile, a fusion with fixed rater expertise might be easier to learn since the static rater preference can be modeled by neural network [1]. However, DF-GT is fused by dynamically generated expertise maps.

In other words, the expertness maps of each sample are different depends on the diagnosis prediction. The standard segmentation network that lacks dynamic adaptability may become hard to follow.

Toward the efficient learning of DF-GT, we propose Take and Give Model (T&G Model) to instill the diagnosis features into the segmentation network dynamically. The architecture of T&G Model is shown in Figure 4 (c). In particular, T&G Model contains a Take Module and a Give Module. Give Module bridges the segmentation encoder and the diagnosis network to fill the segmentation information into the diagnosis features. Through the attention mechanism, Give Module weights the diagnosis feature based on its affinity with the segmentation feature. Only the diagnosis features closely related to the segmentation will be selected and transformed, resulting in a transformed diagnosis feature that can be used for segmentation. Take Module bridges the transformed diagnosis feature and the segmentation decoder to provide the diagnosis knowledge to the segmentation. Take Module is symmetrical to Give Module but picks and transforms the segmentation feature based on the diagnosis features. Take Module and Give Module will be applied on several corresponding layers of the segmentation encoder, the diagnosis network, and the segmentation decoder.

Specifically, T&G Model is constructed by attention mechanism [31] following MLP layer. Consider T&G Model at the \( k \)th layer. The inputs of Give Module are a segmentation feature map and a diagnosis feature map, which are the \( k \)th layers of segmentation encoder and diagnose network, respectively. The output of the module is a transformed diagnosis feature map with the same size as the input feature map. To deal with the feature map efficiently, we reshape the feature map \( F \in \mathbb{R}^{H \times W \times C} \) into a sequence of flattened patches \( f \in \mathbb{R}^{N \times (P^2 \cdot C)} \), where \((H, W)\) is the resolution of the original feature map, \( C \) is the number of channels, \((P, P)\) is the resolution of the patches, and \( N = HW/P^2 \) is the resulting number of patches. Consider the encoder segmentation feature map patches is \( f_{se}^k \) and the diagnosis feature map patches is \( f_{sd}^k \), then Give Module can be represented as:

\[
\hat{f}_d^k = \text{MLP}(\text{Attention}(f_{se}^k + E_{se}^k, f_{sd}^k + E_{sd}^k, f_{d}^k)),
\]

where \( \hat{f}_d^k \in \mathbb{R}^{N \times (P^2 \cdot C)} \) is the transformed diagnosis feature, \( \text{Attention}(\text{query, key, value}) \) denotes multi-head attention mechanism, MLP is multi-layer perceptron, \( E_{se}^k, E_{sd}^k \in \mathbb{R}^{N \times (P^2 \cdot C)} \) are positional encodings [32] for the segmentation feature map and diagnosis feature map, respectively. A standard attention mechanism is adopted here, which first calculates an affinity weight map \( a \in \mathbb{R}^{N \times (H \times W)} \) between the query and key, and then uses it to select the features in value.

The affinity weight map \( a \) is represented as:

\[
a = \text{softmax}(q(f_{se}^k + E_{se}^k) k(f_{sd}^k + E_{sd}^k)^T) / \sqrt{P^2 \cdot C}, \tag{9}
\]

where the functions \( q(\cdot), k(\cdot) \) denote the linear mappings for the inputs of query and key. We can see the normalized affinity weights \( a \) are defined by how all the segmentation features influence each diagnosis feature. These weights are then applied to all the diagnosis features in value,

\[
\text{Attention}(f_{se}^k + E_{se}^k, f_{sd}^k + E_{sd}^k, f_{d}^k) = a \cdot v(f_{d}^k), \tag{10}
\]

where \( v(\cdot) \) is the linear mapping for value.

We then apply a Multi-layer Perceptron (MLP) to further select the reinforced attention result. MLP has two linear mappings with weight \( W_{f1}, W_{f2} \in \mathbb{R}^{(P^2 \cdot C) \times (P^2 \cdot C)} \) and a GELU [33] activation function in its module:

\[
\text{MLP}(f) = \text{GELU}(f \cdot W_{f1}) \cdot W_{f2}. \tag{11}
\]

Take Module provides the segmentation decoder with the diagnosis knowledge to get a diagnosis-focused segmentation result. Take Module is an attention-based module symmetrical to Give Module, which conversely uses combined feature \( \hat{f}_d^k \) as query and segmentation feature \( f_{sd}^k \) as key and value, so as to select and transform the segmentation features based on the diagnosis knowledge. Take Module is applied to the features before the standard deconvolution layer in the decoder. Consider the \( k \)th layer in segmentation decoder is \( f_{sd}^{k+1} \), then Take Module interact with decoder by:

\[
f_{sd}^{k+1} = \text{Deconv} (\text{MLP}(\text{Attention}(f_{sd}^k + E_{sd}^k, f_{sd}^k + E_{sd}^k, f_{sd}^k))) \tag{12}
\]

Give Module selects the diagnosis features by the segmentation features in the segmentation encoder, and Take Module enhances the diagnosis related segmentation features in the segmentation decoder. Through this way, the segmentation network is able to capture the diagnosis calibrated segmentation features in DF-GT, so as to efficiently predict DF-GT from the raw images in the inference stage.

III. EXPERIMENT

A. Tasks and Datasets

Extensive experiments are conducted to verify the effectiveness of the proposed framework. We conduct the experiments on three different medical segmentation tasks with data from varied image modalities, including color fundus images, ultrasound images, and dermoscopic images.
1) OD/OC Segmentation & Glaucoma Diagnosis: On the fundus images, we calibrate the OD/OC segmentation via glaucoma diagnosis. The experiments are conducted on REFUGE-2 benchmark [22], which is a publicly available dataset for OD/OC segmentation and glaucoma classification. It contains 5191 ultrasound images from two sources, including 4554 images from TNSCUI [34] and 637 images from DDTI [35]. The images were segmented manually by five graders, following the same annotation protocol, and their annotations were approved by experienced thyroid radiologists. The images are diagnosed as malignant or benign thyroid nodules, where 2881 samples correspond to malignant subjects, and the others correspond to non-malignant subjects. The malignant samples are distributed equally to the training, validation, and test set.

2) Thyroid Nodules Segmentation & Thyroid Cancer Diagnosis: On ultrasound images, we calibrate thyroid nodule segmentation via thyroid cancer diagnosis. The experiments are conducted on the TNMIX benchmark, a publicly available mixed dataset for thyroid nodule segmentation and diagnosis. It contains 1200 color fundus images, including three sets, each with 400 images for training, validation, and testing. Seven glaucoma experts from different organizations labeled the OD/OC contour masks manually for the REFUGE-2 benchmark. 120 samples correspond to malignant subjects, and the others correspond to non-glaucomatous subjects. The glaucomatous subjects are distributed equally to the training, validation, and test set.

3) Skin Lesions Segmentation & Melanoma Diagnosis: On dermoscopic images, we calibrate skin lesions segmentation via melanoma diagnosis. The experiments are conducted on ISIC [36], an open-source dataset of skin lesions segmentation and diagnosis. The images are associated with a ground-truth diagnosis of melanoma and skin lesions masks. 1600 images with multiple skin lesions masks are selected to conduct the experiment. Skin lesions are annotated individually by four recognized skin cancer experts. 312 samples correspond to melanoma subjects, and the others correspond to non-melanoma subjects. The dataset is split into three sets, with 960 images for training, 320 for validation, and 320 for testing. The proportion of malignant samples in each set is made roughly the same.

B. Experimental Setup

1) Implementation Details: To verify the proposed method, we train two baselines for diagnosis and segmentation, denoted as Baseline-Dia and Baseline-Seg, respectively. Baseline-Dia is implemented by a segmentation attentive diagnosis network [17] using ResNet50 [37] as the backbone. In the pre-training stage, we train the network with majority vote segmentation masks on the training set. Baseline-Seg is a standard UNet [38] using ResNet50 as the backbone. In DiFF, we use the frozen pre-trained Baseline-Dia as the diagnosis network connected to T&G Model.

All the experiments are implemented with the PyTorch platform and trained/tested on 4 Tesla P40 GPUs with 24GB of memory. All training and test images are uniformly resized to the dimension of 256 × 256 pixels. DF-GT is trained using Adam optimizer [39] for 125 epochs. DiFF is trained end-to-end using Adam optimizer with a mini-batch of 16 for 80 epochs. The learning rate is always set to 1 × 10^{-4}. Detailed network structures and hyper-parameters can be found in our released code.

2) Evaluation Metric: The diagnosis performance is evaluated by AUC (Area Under the receiver operating characteristic Curve). The segmentation performance is evaluated by soft dice coefficient ($D$) through multiple threshold levels, set as (0.1, 0.3, 0.5, 0.7, 0.9). At each threshold level, the predicted probability map and soft ground-truth are binarized with the given threshold, and then the dice metric [40] is computed. $D$ scores are obtained as the averages of multiple thresholds.

C. Experiment Results

In the experiments, we first verify the effectiveness of the proposed DF-GT and T&G Model by comparing them with corresponding SOTA methods in Section III-C.1 and Section III-C.2, respectively. Then we verify the overall performance of DiFF with previous multi-rater learning methods in Section III-C.3. Finally, we conduct a detailed ablation study to verify the effectiveness of each proposed component in Section III-C.4.

1) DF-GT vs SOTA Labels-Fusion Methods: In this part, we compare DF-GT with SOTA multi-rater fusion strategies. Specifically, we validate the diagnosis performance of the fused segmentation labels by various segmentation-assisted diagnosis approaches. In particular, the multi-rater labels are first fused to unique ground-truth by the multi-rater fusion strategies. Then the unique ground-truth will be sent to a series of segmentation-assisted diagnosis models to evaluate its diagnosis contribution.

In the comparison, the multi-rater fusion strategies include majority vote (MV), STAPLE [8], AggNet [10], and MaxMig [11]. The segmentation-assisted diagnosis models we selected for the evaluation include the concatenation based method (DL-CAT) implemented by ResNet50 backbone [37], ROI based method (DL-ROI) [20], the Attention based method (DL-ATT) [19], the Semi-supervised Learning based method (DL-SSL) [18], and the transformer based feature-fusion method (DL-ScAT) [21].

DL-ROI is a two-stage method. Initially, it utilizes Faster RCNN to extract the bounding box of the region of interest (ROI). In the subsequent stage, this ROI is classified. In our experiments, we supply the segmentation results corresponding to the ROI area to the secondary recognition network.

DL-ATT incorporates an attention subnetwork designed to predict areas that ophthalmologists may concentrate on. In experiments, we concatenate the segmentation results with the attention map produced by this subnetwork.

DL-SSL integrates segmentation results into a teacher-student learning paradigm. In this study, the color fundus images are initially concatenated with their corresponding cup/disc masks, which are then fed into the teacher model.

https://github.com/WuJunde/DiagnosisFirst
we show the visualized comparison in Figure 6. We can see
of diagnosis methods, not only the deep learning based ones.
performance, indicating it is generally effective for all kinds
coma diagnosis, DF-GT also presents the highest diagnosis
other fused ground-truths. On vCDR based models of glau-
contrast, the methods show similar performance on DL-ROI,
are more efficiently used to facilitate the diagnosis. By con-
ATT, DL-SSL, and DL-SeAT, in which segmentation features
proposed ExpG on glaucoma, thyroid cancer and melanoma diagnosis
tasks respectively. We evaluate the diagnosis performance of them by
different diagnosis models, including vCDR-TS [41], vCDR-SVM [42], DL-
cat [23], DL-ROI [20], DL-ATT [19], DL-SSL [18], and DL-SeAT [21]. (d):
Comparison of original high-frequency DF-GT (denoted as Original) and
ExpG generated low-frequency DF-GT (denoted as ExpG) on different
tasks.

DL-SeAT proposes an asymmetric multi-scale interaction
strategy along with a SeA-block strategy to invigorate diagnostic
features by using correlated segmentation features, thereby
improving diagnostic performance.

In glaucoma diagnosis task, directly calculating the vertical
Cup-to-Disc Ratio (vCDR) on OD/OC is also a commonly
used method in the clinical glaucoma diagnosis. Thus we
additionaly use two vCDR based diagnosis methods, vCDR-
TS [41] and vCDR-SVM [42], which implemented by
thresholding and SVM on vCDR, respectively. The quanti-
tative comparison measured by AUC (%) on three different
tasks is shown in Figure 5 (a), (b), and (c).

In Figure 5, we can see DF-GT outperforms the other multi-rater fusion methods on all of the segmentation-assisted
diagnosis models. The improvement is more significant on DL-
ATT, DL-SSL, and DL-SeAT, in which segmentation features
are more efficiently used to facilitate the diagnosis. By con-
trast, the methods show similar performance on DL-ROI,
which used segmentation only for localization. It demonstrates
that DF-GT has a much higher diagnostic value than the
other fused ground-truths. On vCDR based models of glau-
coma diagnosis, DF-GT also presents the highest diagnosis
performance, indicating it is generally effective for all kinds
of diagnosis methods, not only the deep learning based ones.

For a detailed comparison with the multi-rater fusion labels,
we show the visualized comparison in Figure 6. We can see
that MV and STAPLE highlight the regions where the multiple
raters agreed and play down the uncertain regions. However,
it has the chance that the minority is correct and lead the
discriminative features to be obliterated. The negative effects
of these examples constrain their diagnosis performance.

AggNet and MaxMig depend on the segmentation prediction
to fuse the labels. Thus they show inferior visualized and
diagnosis performance on the tasks that the contours of the
lesions/organs are more ambiguous, i.e., glaucoma and thyroid
nodule. By contrast, the proposed DF-GT is able to pick and
reserves the segmentation features that facilitate the diagnosis
and thus achieves better performance. For example, on the skin
lesion segmentation in Figure 6, we can see DF-GT uniquely
stresses the asymmetrical and irregular structure of the lesion,
which are the key signs of melanoma. On glaucoma diagnosis,
DF-GT fuses the labels with larger vCDR, which is a crucial
biomarker indicating glaucoma.

Figure 7 shows the visualized examples of all high-
frequency elimination methods. The first line is a non-
glaucoma example, and the second is a glaucoma example.
In the figure, “High” refers to the segmentation ground truth
obtained without any high-frequency elimination methods;
“Activation”, on the other hand, refers to the segmentation
ground truth acquired after applying the proposed ExpG
method, which is then overlaid on the input image. We can
see TransRob and Fourier’s visualized effects are obviously
improved compared with high-frequency examples. But ExpG
achieves much better visualization effects than both of them.
In the experiments, we adopt ExpG to eliminate the high-
frequency components in DF-GT optimization process.

Figure 8 shows the annotations produced by multiple raters
and their expertness maps obtained with and without ExpG,
where blue indicates lower weights and red signifies higher
weights. It can be observable that without ExpG, the expert-
ness maps tend to fit the fixed-parameter diagnostic network
due to back propagation. Therefore, the expertness maps
of different raters all contain a lot of high-frequency information,
and mostly of high-frequency information are similar. For exam-
ple, there are some similar contours in the expertness maps at
the edge of the annotation given by each rater. However, upon
introducing the high-frequency elimination method, ExpG,
a consistently lower amount of high-frequency information is
present in the expertness map.

2) T&G Model Vs SOTA Segmentation Methods: In this
part, we compare T&G Model with previous SOTA segment-
mentation methods on three different diseases. Since we find
the segmentation models proposed for the specific diseases
usually perform better than the general ones, we select the SOTA models proposed for each disease respectively for the comparison. On OD/OC segmentation, we compare with CENet [43], AGNet [44], BEAL [45], pOSAL [46], and ATTNNet [47]. On thyroid nodule segmentation, we compare with TNSCNN [48], CasCNN [49], MPCNN [50], TRFENet [51], and TNSNet [52]. On skin lesion segmentation, we compare with DDN [53], JaccNet [54], Unver [55], FrCN [56], and iFCN [57]. Segmentation performance measured by Dice Score is shown in Table I.

In Table I, we can see the proposed model consistently achieves superior segmentation performance compared with SOTA segmentation methods, which indicates T&G Model can better predict DF-GT from raw images. The improvement is especially significant when the object is ambiguous. For example, the proposed T&G Model outperforms the previous SOTA by a remarkable 2.96% on the optic cup, which is quite hard to discriminate from the raw images. It indicates that T&G Model can select and enhance the unique features conducive to the diagnosis rather than only using the structural features on the raw images like most other segmentation methods.

3) DIFF vs SOTA Multi-Rater Learning Methods: In this section, we show DIFF, a combination of DF-GT and T&G Module, can predict the segmentation maps that are more conducive to the diagnosis from the raw images. Toward that end, we conduct an overall comparison between DIFF and the previous multi-rater learning methods. The compared methods include the Bayesian based methods: sPU-Net [7] and PHIseg [58], the multi-head based method: WDNet [5], the Expectation-Maximization (EM) based methods: AggNet [10] and MaxMig [11], and calibrated segmentation method: MRNet [1]. We quantitatively evaluate the predicted segmentation results against each of the multi-rater labels and the self-fused labels. Self-fused labels refer to the weighted sum of the multi-rater labels by the self-predicted expertness, as shown in Eqn. (3). Since in MRNet and WDNet, the expertness is user-provided, we evaluate their performance under the majority vote, following the setting of their papers. In sPU-Net and PHIseg, we sample 16 times from the latent space following their setting and take the average as the result.

The segmentation capability of the predicted segmentation maps on multi-rater (R1-R5) annotations and self-fusion labels is shown in Table II.

Comparing the diagnosis capabilities of the methods, we can see DIFF consistently achieves superior performance and outperforms the second best method by 2.73%, 2.51%, and 2.04% AUC on glaucoma, thyroid cancer, and melanoma diagnosis, respectively. This indicates that the proposed DIFF can provide more effective results in scenarios concerned with the final diagnostic effect. Comparing the segmentation performance

| Methods      | R1     | R2     | R3     | R4     | R5     | Self   | Diagnosis |
|--------------|--------|--------|--------|--------|--------|--------|-----------|
| sPU-Net      | 81.43  | 82.54  | 81.75  | 83.43  | 83.76  | 84.29  | 82.73     |
| PHIseg       | 80.82  | 80.37  | 83.29  | 84.20  | 81.45  | 83.67  | 84.68     |
| WDNet        | 76.32  | 77.43  | 79.59  | 78.45  | 77.21  | 80.38  | 80.04     |
| AggNet       | 79.37  | 80.32  | 84.79  | 83.11  | 81.04  | 84.85  | 83.38     |
| MaxMig       | 82.21  | 81.04  | 82.62  | 83.54  | 81.66  | 85.10  | 82.90     |
| MRNet        | 80.92  | 81.31  | 84.20  | 85.42  | 82.06  | 84.72  | 85.24     |
| DIFF         | 82.18  | 84.53  | 86.21  | 85.89  | 87.31  | 87.75  | 87.75     |

In Table III, we compare the proposed DIFF and SOTA multi-rater learning algorithms on Thyroid Nodules Segmentation & Thyroid Cancer Diagnosis. We evaluate the segmentation capability of the predicted segmentation maps on multi-rater (R1-R5) annotations and self-fusion labels. We evaluate the diagnosis capability of the predicted segmentation maps on Baseline-DIA. Best results are shown in bold.

| Methods      | R1     | R2     | R3     | R4     | R5     | Self   | Diagnosis |
|--------------|--------|--------|--------|--------|--------|--------|-----------|
| sPU-Net      | 81.43  | 82.54  | 81.75  | 83.43  | 83.76  | 84.29  | 82.73     |
| PHIseg       | 80.82  | 80.37  | 83.29  | 84.20  | 81.45  | 83.67  | 84.68     |
| WDNet        | 76.32  | 77.43  | 79.59  | 78.45  | 77.21  | 80.38  | 80.04     |
| AggNet       | 79.37  | 80.32  | 84.79  | 83.11  | 81.04  | 84.85  | 83.38     |
| MaxMig       | 82.21  | 81.04  | 82.62  | 83.54  | 81.66  | 85.10  | 82.90     |
| MRNet        | 80.92  | 81.31  | 84.20  | 85.42  | 82.06  | 84.72  | 85.24     |
| DIFF         | 82.18  | 84.53  | 86.21  | 85.89  | 87.31  | 87.75  | 87.75     |
on multi-rater labels, DiFF achieves top-3 on almost all of the raters, but its distribution over the raters is not always consistent with the other methods.

For example, on glaucoma diagnosis, most other methods prefer R1, R2, and R6 among the seven raters, while DiFF prefers R1, R2, and R4. This discrepancy may come from the criteria used to evaluate the rater’s expertise. Most previous methods leverage the raw image prior to evaluate the rater expertise, which means that the annotations with a structure more similar to that observed on the raw image will be given higher confidence. By contrast, DiFF leverages the diagnosis as the standard to evaluate the raters. As shown in Figure 3, R4 is more beneficial to the diagnosis than R6. Thus the inclination toward R4 of DiFF demonstrates that it is able to recognize the diagnosis-focused raters and predict the corresponding features directly from the raw image.

It is also worth noting that DiFF may also show consistent inclinations toward the raters with the other methods. For example, on thyroid cancer diagnosis, all these methods incline to R2, R3, and R4. The different point is DiFF shows a heavier inclination and thus wins the highest diagnosis score. It seems to indicate that those annotations are more accurate and a more precise annotation of the structures on the raw images can indeed bring higher diagnosis performance in some cases. However, when the segmentation performance has reached a certain level and the boundary of the target object is really ambiguous, like many of the optic cups on fundus images, simply following the raw image structure will not be effective anymore for the diagnosis. In such cases, some of the raters are likely to annotate the objects with their prior knowledge of disease diagnosis. For example, when the contours of OD/OC are ambiguous, the rater may deliberately annotate OD/OC with larger vCDR if he/she suspects the case is positive (glaucoma). In these cases, DiFF is able to get the message and enhance these features through the diagnosis-first principle.

### 4) Ablation Study: DF-GT

ExpG is proposed in the paper to improve the generalization of DF-GT. To verify the effectiveness of ExpG, we compare the DF-GT before and after the application of ExpG. We use the same evaluation setting as that of multi-rater labels fusion strategies in Section III-C.1.

The diagnosis performance measured by AUC (%) is shown in Figure 5 (d).

As shown in Figure 5 (d), compared with original high-frequency DF-GT, ExpG-generated DF-GT consistently achieves higher performance on different diagnosis methods, especially on stronger networks, like DL-ATT, DL-SSL, and DL-SeAT. Concretely, ExpG-generated DF-GT outperforms the counterpart by a 5.93% AUC on DL-ATT, a 7.90% AUC on DL-SSL, a 7.59% AUC on DL-SeAT on glaucoma diagnosis. Similar conclusions can also be drawn from the other tasks. It clearly shows that ExpG-generated DF-GT is more robust for different diagnosis networks.

### T&G Model

We propose T&G Model to segment DF-GT from raw images. In order to verify the effectiveness of T&G Model, we compare the segmentation performance of Baseline-Seg on DF-GT before and after the application of T&G Model. The quantitative results on OD/OC segmentation are shown in Table V. We can see that compared with the segmentation baseline without T&G Model, T&G Model equipped one shows a significant improvement on DF-GT. That is because the segmentation network that absents the prior diagnosis knowledge is hard to capture the variance of DF-GT from the raw images. The proposed T&G Model instills diagnosis knowledge in the segmentation network, thus improving the segmentation performance on diagnosis-first labels.

The selection of the connected layers of T&G Model will also affect the final segmentation performance. To find which layers should be connected by T&G Model, the ablation study is performed in Table VI. We sequentially apply the proposed T&G Model on four different residual blocks of the segmentation encoder, diagnosis network, and segmentation decoder. We denote the residual blocks from shallow to deep as B1 to B4. As shown in the table, applying T&G Model on the shallow layers seems to be more efficient than applying it on the deeper layers. This may be because the diagnosis feature in B4 may be too unique to be leveraged for the segmentation. The segmentation network achieves the best performance when connecting B1, B2, and B3, but the effect of connecting the middle two blocks (B2 and B3) is the most obvious. This indicates that the diagnosis features

---

**Table IV**

| Methods | R1 | R2 | R3 | R4 | Self | Diagnosis |
|---------|----|----|----|----|------|-----------|
| SPU-Net ([7]) | 87.78 | 85.69 | 81.63 | 86.37 | 87.79 | 80.15 |
| PHiSeg ([58]) | 86.26 | 84.21 | 82.02 | 85.71 | 87.50 | 79.43 |
| WDNNet ([5]) | 79.64 | 78.35 | 80.40 | 78.29 | 81.20 | 76.66 |
| AggNet ([10]) | 83.98 | 81.12 | 82.78 | 84.79 | 85.85 | 78.52 |
| MaxMig ([11]) | 87.53 | 85.75 | 85.13 | 86.45 | 87.44 | 80.17 |
| MRNet ([13]) | 87.99 | 86.76 | 84.71 | 85.69 | 88.02 | 80.32 |
| DiFF | **88.81** | **87.97** | **83.21** | **84.70** | **89.48** | **82.06** |

---

**Table V**

| Connected Blocks | Ave Dice |
|------------------|----------|
| B1 | 95.30 | 86.61 |
| B2 | 95.38 | 84.72 |
| B3 | 94.79 | 85.34 |

---

**Table VI**

| Connected Blocks | Ave Dice |
|------------------|----------|
| B1 | 95.30 | 86.61 |
| B2 | 95.38 | 84.72 |
| B3 | 94.79 | 85.34 |
| B4 | 93.36 | 82.78 |

---

**Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.**
nostic performance, as shown in Figure 9. Our results indicate
raters in three task settings and evaluated the impact on diag-
map decreases notably. To assess this, we varied the number of
vary significantly, the diagnostic accuracy of the segmentation
When fewer raters are involved, especially if their assessments
improvement and attention.

D. Limitations and Failure Cases

In this part, we discuss some of the limitations and potential failure cases of our method, shedding light on areas for further improvement and attention.

One key limitation is the sensitivity of our final diagnosis performance to the number of raters used during training. When fewer raters are involved, especially if their assessments vary significantly, the diagnostic accuracy of the segmentation map decreases notably. To assess this, we varied the number of raters in three task settings and evaluated the impact on diagnostic performance, as shown in Figure 9. Our results indicate a clear disparity in performance: with fewer raters, there’s a noticeable drop in diagnostic accuracy, as measured by the AUC. Performance generally increases linearly with up to five raters, then begins to plateau. This is likely because fewer raters may overlook critical regions key to accurate diagnosis, challenging our method to create an effective segmentation from the limited input.

Another limitation of our method is its reduced performance in segmenting blurred or over-exposed images. Under such conditions, our approach may produce artifacts, blurred segments, or fail to predict the correct segmentation. We can refer to these images as degraded or low-quality images, as there are fewer in the database and the proposed method has not been optimized for such images, resulting in poor processing effects. This issue is exemplified in the optic-cup segmentation on the REFUGE dataset, as illustrated in Figure 10. When organ boundaries in raw images are unclear, our method struggles to correctly assign diagnostic prior to the segmentation network, leading to unsatisfied results.

In summary, although our method demonstrates promising results, the highlighted limitations and failure cases emphasize the need for ongoing research and enhancement.

To address the issue of insufficient raters, one potential approach is to generate more diverse pseudo masks using the generative models [59]. Regarding the segmentation failures with low-quality images, preprocessing the images with deblurring [60] or overexposure correction methods [61], [62] could be beneficial before they enter our pipeline. We plan to explore and refine these solutions in our future work, aiming to enhance the robustness of our framework.

Despite these challenges, we believe our method represents a significant contribution to the field and provides a solid foundation for future improvements.

IV. Conclusion

In this work, we proposed calibrating inter-observer segmentation uncertainty via the diagnosis-first principle. In practice, we designed DiFF, a framework that can segment the masks to improve the diagnosis performance. This was achieved by the use of DF-GT, a fusion of multi-rater segmentation labels in favor of the diagnosis, as well as T&G Model equipped segmentation network to learn from DF-GT. Extensive empirical experiments demonstrated the overall superior diagnosis performance of DiFF. In future work, we will explore the relationship between the diagnosis-first segmentation features and the clinical biomarkers to explain how neural networks leverage these features to make the diagnosis decision. We will also explore how humans visualize and analyze these diagnosis-first segmentation features.

References

[1] W. Ji et al., “Learning calibrated medical image segmentation via multi-rater agreement modeling,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2021, pp. 12341–12351.

[2] S. Yu et al., “Difficulty-aware glaucoma classification with multi-rater consensus modeling,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervent., 2020, pp. 741–750.

[3] A. W. Callhoun, E. A. Rider, E. C. Meyer, G. Lamiania, and R. D. Truong, “Assessment of communication skills and self-appraisal in the simulated environment: Feasibility of multirater feedback with gap analysis,” Simul. Healthcare, J. Soc. Simul. Healthcare, vol. 4, no. 1, pp. 22–29, 2009.

[4] H. Fu et al., “A retrospective comparison of deep learning to manual annotations for optic disc and optic cup segmentation in fundus photographs,” Transl. Vis. Sci. Technol., vol. 9, no. 2, p. 33, Jun. 2020.

[5] M. Y. Guan, V. Gulshan, A. M. Dai, and G. E. Hinton, “Who said what: Modeling individual labelers improves classification,” in Proc. 32nd AAAI Conf. Artif. Intell., 2018, pp. 3109–3118.

[6] H. Chou and C. Lee, “Every rating matters: Joint learning of subjective labels and individual annotators for speech emotion classification,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 5886–5890.

[7] S. Kohl et al., “A probabilistic U-Net for segmentation of ambiguous images,” in Proc. Adv. Neural Inf. Process. Syst., vol. 31, 2018, pp. 6965–6975.

[8] S. K. Warfield, K. H. Zou, and W. M. Wells, “Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation,” IEEE Trans. Med. Imag., vol. 23, no. 7, pp. 903–921, Jul. 2004.

[9] V. C. Raykar et al., “Learning from crowds,” J. Mach. Learn. Res., vol. 11, no. 4, pp. 1297–1322, 2010.

[10] S. Albarqouni, C. Baur, F. Achilles, V. Belagianannis, S. Demirci, and N. Navab, “AggNet: Deep learning from crowds for mitosis detection in breast cancer histology images,” IEEE Trans. Med. Imag., vol. 35, no. 5, pp. 1313–1321, May 2016.
