Learning Calibrated Medical Image Segmentation via Multi-rater Agreement Modeling

Wei Ji[1,2], Shuang Yu[1]✉, Junde Wu[1], Kai Ma[1], Cheng Bian[1], Qi Bi[1]
Jingjing Li[2], Hanruo Liu[3], Li Cheng[2]✉, Yefeng Zheng[1]
[1]Tencent Jarvis Lab, Shenzhen, China [2]University of Alberta, Canada
[3]Beijing Tongren Hospital, Capital Medical University, Beijing, China
{wji3, lcheng5}@ualberta.ca, {shirlyyu, kylekma, yefengzheng}@tencent.com

Abstract

In medical image analysis, it is typical to collect multiple annotations, each from a different clinical expert or rater, in the expectation that possible diagnostic errors could be mitigated. Meanwhile, from the computer vision practitioner viewpoint, it has been a common practice to adopt the ground-truth labels obtained via either the majority-vote or simply one annotation from a preferred rater. This process, however, tends to overlook the rich information of agreement or disagreement ingrained in the raw multi-rater annotations. To address this issue, we propose to explicitly model the multi-rater (dis-)agreement, dubbed MRNet, which has two main contributions. First, an expertise-aware inferring module or EIM is devised to embed the expertise level of individual raters as prior knowledge, to form high-level semantic features. Second, our approach is capable of reconstructing multi-rater gradings from coarse predictions, with the multi-rater (dis-)agreement cues being further exploited to improve the segmentation performance. To our knowledge, our work is the first in producing calibrated predictions under different expertise levels for medical image segmentation. Extensive empirical experiments are conducted across five medical segmentation tasks of diverse imaging modalities. In these experiments, superior performance of our MRNet is observed comparing to the state-of-the-arts, indicating the effectiveness and applicability of our MRNet toward a wide range of medical segmentation tasks. Source code is publicly available.

1. Introduction

Accurate anatomy and lesion segmentation is crucial in clinical assessment of various diseases, including for example glaucoma [28, 36, 43], prostate diseases [30, 52], and brain tumors [11, 17, 44]. It has been increasingly popular to develop automated segmentation systems, to facilitate a reliable reference for the quantification of disease progression, which is especially accelerated by the exciting breakthroughs of deep convolutional neural networks (CNNs) [7, 20, 34, 35, 49, 55, 56, 59] over the past decade.

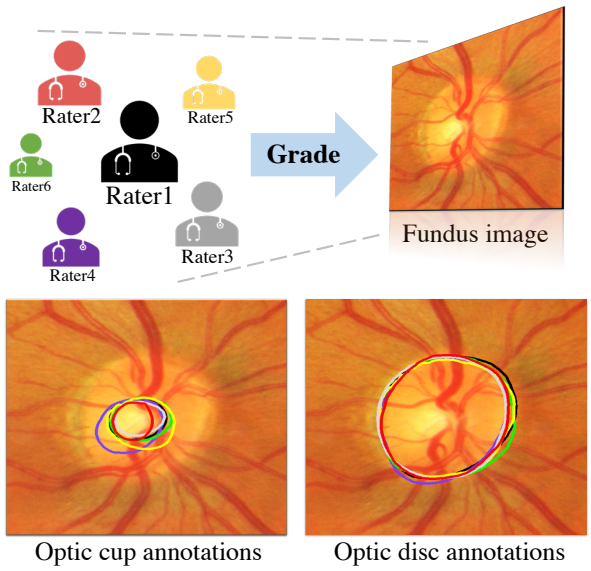

Figure 1. **Top:** an exemplar medical image grading scenario conducted by multiple raters with different expertise levels. **Bottom:** visualization of optic cup and disc annotations of the above raters.

Different from labelling natural images, medical images are often independently annotated by *a group of experts or raters*, to mitigate the subjective bias of a particular rater due to factors such as the level of expertise, or possible negligence of subtle symptoms [13, 39, 23, 28]. Inter-observer variability, as frequently reported by relevant research in the clinical field, often leads to challenges in segmenting highly uncertain regions [3, 23, 37]. Fig. 1 provides a representative illustration of the multi-rater grading process in

Wei Ji, Shuang Yu and Junde Wu have equal contributions. Wei Ji contributes to this work during internship at Tencent Jarvis Lab.
Shuang Yu and Li Cheng are the corresponding authors.

annotating optic cups and discs from fundus images, with notable uncertainties or disputed regions presented among graders. It is thus necessary for automated systems to consider a proper segmentation strategy that reflects the underlying (dis-)agreement among multiple experts. Existing works typically require unique ground-truth annotations, each pairing with one of the input images to train the deep learning models. It is a common practice to take majority vote, STAPLE [50] or other label fusion strategies to obtain the ground-truth labels [5, 29, 30, 34, 57, 59]. Being simple and easy to implement, this strategy, however, comes at the cost of ignoring altogether the underlying uncertainty information among multiple experts. Very recently, several efforts start to explore the influence of multi-rater labels by label sampling [19, 24] or multi-head [16] strategies. It is reported that models trained with multi-rater labels are better calibrated than those with the typical ground-truth label via, e.g. majority vote, which are prone to be over-confident [19, 24].

Meanwhile, there still lacks a principled approach to incorporate in training the rich uncertainty information from multiple raters. Specifically, we focus on the following questions: 1) how to integrate varied expertise-level, or *expertness*, of individual raters into the network architecture? 2) how to exploit the uncertainty information among different experts to produce probability maps that better reflect the underlying graders' (dis-)agreement? This inspires us to propose a multi-rater agreement modeling framework, MRNet. To our knowledge, it is the first in explicitly addressing the above-mentioned questions. Our framework has the following three main contributions:

- The notion of *expertness* is explicitly introduced as prior knowledge about the expertise levels of the involved multi-raters. It is embedded in the high-level semantic features through the proposed Expertise-aware Inferring Module (EIM), enabling the representation capability to accommodate the multi-rater settings.
- A Multi-rater Reconstruction Module (MRM) is designed to reconstruct the raw multi-rater gradings from the the expertness prior and the soft prediction of the model. This enables the estimation of an uncertainty map that reflects the inter-rater variability, by exploiting the intrinsic correlations between the fused soft label and the raw multi-rater annotations.
- To better utilize the rich cues among multi-rater (dis-)agreements, we further incorporate in our framework a Multi-rater Perception Module (MPM), which empirically leads to noticeable performance boost.

Extensive experiments are performed on five different medical image segmentation tasks of diverse image modalities, including color fundus imaging, computed tomography (CT), and magnetic resonance imaging (MRI). Overall, our MRNet framework consistently outperforms the state-of-the-art methods as well as existing multi-rater strategies. In addition, our MRNet runs in real-time (29 frame per second) at inference stage, making it practically appealing for many real-world applications.

2. Related Work

Medical Image Segmentation. With the advancement of CNNs, an increasing number of deep learning architectures have been proposed for medical segmentation tasks such as optic disc/cup segmentation [60, 29, 57, 12] in fundus images, prostate segmentation [21, 30, 48] and brain tumor segmentation [4, 6]. These methods have obtained superior performance comparing to traditional feature engineering based methods [8, 9, 10]. Taking optic disc/cup segmentation as an example, Fu et al. [12] proposed a U-shaped network with multi-scale supervision strategy for polar transformed fundus images to produce the segmentation maps. Gu et al. [15] integrated dense atrous convolution block and residual multi-kernel pooling to U-Net structure to capture high-level features with context information. Zhang et al. [58] presented an attention guided network using guided filter to preserve the structural information and reduce the negative influence of background. Meanwhile, Li et al. [29] integrated detection and multi-class segmentation into a unified architecture for segmenting the optic cup and disc regions. Wang et al. [45] attempted to utilize the designed domain adaptation frameworks for fundus image segmentation, in order to increase the cross-domain prediction accuracy.

A common practice adopted by the above-mentioned methods, as well as most existing CNNs based learning methods, is to construct training examples by retaining unique ground-truth labels for each of the training instances. In this manner, the valuable multi-rater labels obtained in the grading procedure with inter-rater variability are unfortunately not well-exploited.

Multi-rater Strategies. Very recently, the problems of the multi-rater labels and inter-rater variability start to attract research attentions [16, 19, 2, 24, 42, 54]. Jensen et al. [19] adopted a label sampling strategy for skin disease classification, by sampling labels randomly from the multi-rater labeling pool during each training iteration. It was observed that model trained with the traditional unique ground-truths would be over-confident, meanwhile model trained with label sampling strategy was better calibrated. Similar observation was also reported by [24] for segmentation task. Label sampling strategy was also utilized by [26] to train a probabilistic model based on a combined U-Net with conditional variational autoencoder to obtain multiple plausible hypotheses. Similarly, Baumgartner et al. [2] employed label sampling strategy as well to train the hierarchical probabilistic model with multi-scale latent variables when using

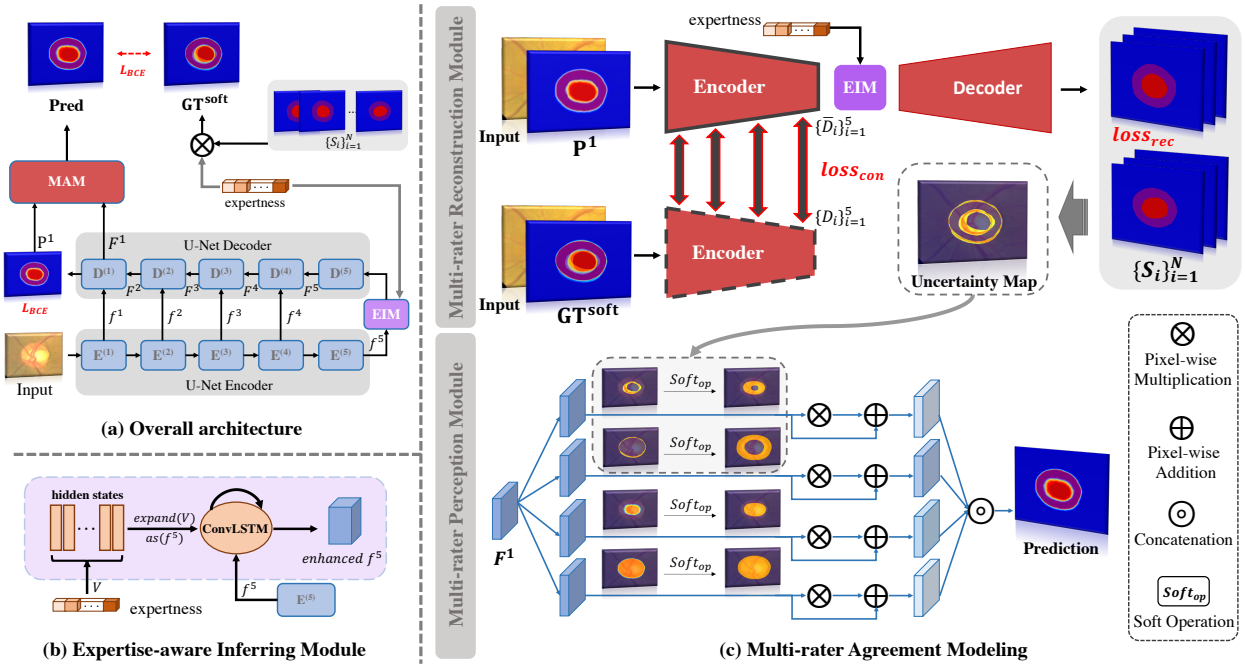

Figure 2. An illustration of our MRNet framework, which starts from (a) an overview of the processing pipeline, and continues with zoomed-in diagrams of individual modules, including (b) the Expertise-aware Inferring Module (EIM), and (c) the Multi-rater Agreement Modeling (MAM) that consists of the Multi-rater Reconstruction Module (MRM), and the Multi-rater Perception Module (MPM).

labels from multiple annotators. Guan et al. [16] predicted the gradings of each rater individually and learned the corresponding weights for final prediction. Yu et al. [54] proposed a multi-branch structure to generate three predictions under different sensitivity settings, to leverage multi-rater consensus information for glaucoma classification.

With the existing multi-rater strategies of label sampling [19, 2, 24] and multiple-head/branch architecture [16, 54], there still lacks a principled research investigation on exploiting the rich (dis-)agreement information among raters in model training and predictions.

3. Methodology

3.1. Motivation

As aforementioned, the inter-grader variability is a well-known issue in the medical image annotation process, since experts differ from each other in their grading preferences and levels of expertise [13, 39, 23, 28]. In order to quantitatively demonstrate such difference, a preliminary experiment is performed with an optic cup segmentation setting on the RIGA benchmark dataset [1].

We train a U-Net [38] using individual rater's annotations for the optic cup segmentation task, and thus obtain six different models (named Model 1-6) corresponding to six raters (named Rater 1-6). The performance of each model against each rater's grading as well as the final consensus label from majority vote is listed in Table 1. It is obvious that all the models have the optimal performance when trained and evaluated with the same rater's annotations but much worse when evaluated by others' annotations. Moreover, when evaluated with the consensus labels obtained with majority vote, Model 1 achieves the best result, followed by Model 2, which is consistent with the database and grader analysis reported in [1]. Two findings can therefore be drawn from here and possibly generalized to medical analysis tasks beyond optic cup segmentation: **1)** individual expert has specific and consistent grading patterns and **2)** the expertise levels among a group of graders are usually different from one to the other. This preliminary study and subsequent findings motivate us to propose our MRNet framework to be discussed next.

Table 1. A preliminary test in examining the grading consistency and expertise level of individual raters, conducted for the optic cup segmentation task on RIGA test set [1] (measured by Dice coefficient). Models 1-6 denote the U-Net models supervised by individual rater's grading. The Raters 1-6 and Majority Vote indicate the labels based on which the model performance is evaluated.

	Rater1	Rater2	Rater3	Rater4	Rater5	Rater6	Majority Vote
Model1	**0.852**	0.823	0.815	0.832	0.795	0.755	**0.866**
Model2	0.834	**0.836**	0.785	0.823	0.784	0.764	0.854
Model3	0.829	0.800	**0.833**	0.786	0.813	0.765	0.851
Model4	0.798	0.809	0.770	**0.875**	0.725	0.691	0.818
Model5	0.803	0.775	0.790	0.731	**0.817**	0.774	0.817
Model6	0.790	0.764	0.763	0.704	0.799	**0.803**	0.797

3.2. Overall Framework

In this work, we propose a novel medical image segmentation framework, named as MRNet, that takes under-

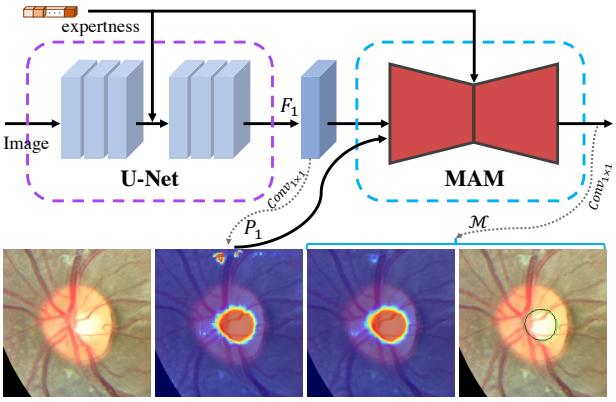

Figure 3. Intermediate visual results in the processing pipeline of our MRNet framework. (a) Input fundus image. (b) Heat map of the initial cup prediction P^1. (c) Heat map of the final refined cup prediction $\mathcal{M}$. (d) Segmentation boundary of the cup prediction (green) and ground-truth (black).

lying agreement/disagreement information among multiple raters into consideration. Fig. 2 illustrates the overall framework of the proposed MRNet, which contains a coarse to fine two-stage processing pipeline. The first stage adopts the widely used U-Net architecture [38] with a ResNet34 [18] backbone pretrained from ImageNet as the encoder part. Then an Expertise-aware Inferring Module (EIM) is inserted at the bottleneck layer to embed the expertise information of individual raters, named as *expertness* vector, into the extracted high-level semantic features of the network. The enhanced feature f^5 is further passed to the decoder blocks of U-Net to generate multi-level decoder features $\{F^i\}_{i=1}^5$. The final decoded feature F^1 is processed by a 1×1 convolutional operation followed by a sigmoid activation function to generate the coarse prediction P^1.

The second stage, aiming to refine the coarse prediction results from the first stage to get better predictions, is composed of two modules arranged in a sequential order. The Multi-rater Reconstruction Module (MRM) is designed to reconstruct the raw multi-rater's gradings, based on which to estimate the pixel-wise uncertainty map that represents the inter-observer variability across different regions. Furthermore, the Multi-rater Perception Module (MPM) with soft attention mechanism is proposed to utilize the uncertainty map to refine the coarse prediction. For simplicity, we use Multi-rater Agreement Modeling (MAM) to represent the combination of the two sequential modules. A simplified illustration of the pipeline with intermediate results is also shown in Fig. 3.

3.3. Expertise-aware Inferring Module

Considering that different experts have different levels of clinical expertise and thus should be assigned with different weights during the model training procedure, we propose an Expertise-aware Inferring Module (EIM) to take advantage of the expertise levels of individual raters as prior knowledge, which is embedded into the segmentation network in the format of conditional information to increase the dynamical representation capability of the extracted features.

In the EIM module, the expertise level cues of multiple raters are formed as a normalized *expertness* vector $V \in \mathbb{R}^{1\times1\times N}$, where N represents the total number of raters and $\sum_{i=1}^{N} V_i = 1$. It is fed to the network as prior knowledge and determines the actual soft GT labels that are set as the network's target. Specifically, the soft GT label used in the training is determined by the annotations of individual raters multiplied by their corresponding weight in the *expertness* vector V, which is denoted as:

$$GT^{soft} = \sum_{i=1}^{N} S_i V_i \rightarrow \varphi(x, V), \tag{1}$$

where φ denotes the model parameters; x is the input image; and S_i means the annotation mask by the i^{th} expert.

During each training iteration, the *expertness* vector V is dynamically set with three different strategies alternatively, including the majority vote mode (i.e., uniform weight among all raters), single rater mode (i.e., assign weight of 1 to single random rater and suppress the rest raters to 0), and random weight assignment (i.e., assign each rater's weight randomly and then normalize to a unit vector). By using different strategies to assign the *expertness* vector, the model learns to associate the influence/weight of individual raters on the final soft predictions. In addition, the dynamic *expertness* vector together with the adaptively changing GT label works as an effective data augmentation strategy that increases the data variability and input-output data pairs being fed to the model. In the inference stage, only the majority vote mode is used by default to set the *expertness* vector, making it easily applicable for clinical applications.

In order to integrate the multi-rater expertise cues into the semantic feature representation effectively, we utilize a ConvLSTM module [40] to generate the enhanced features embedded with the *expertness* vector as hidden state, as shown in Fig. 2(b). ConvLSTM is a powerful recurrent model that not only captures the correlation between features and different expertise levels (i.e., the hidden state), but also summarizes the discriminative dynamic features. To be more specific, we take the feature map from the bottleneck layer (i.e., f^5) as input to the proposed EIM and use the normalized *expertness* vector $V \in \mathbb{R}^{1\times1\times N}$ as initial hidden state h_0. To transfer the *expertness* vector into a proper format for ConvLSTM, we expand V to the same dimension as that of f^5. The procedure can be defined as:

$$h_t = \overset{t}{\circlearrowleft} \text{ConvLSTM}(f^5, h_{t-1}), t = 1, 2, ..., T, \tag{2}$$

where t denotes the time step in ConvLSTM and $\overset{t}{\circlearrowleft}$ indicates the iteration process at time t. After T steps, which

is empirically set as two in this work, an enhanced feature $f^{5e} = h_T$ embedded with expertise cues is generated. The enhanced f^{5e} is further sent to the U-Net decoder to obtain the coarse calibrated prediction P^1 and decoded feature F^1.

3.4. Multi-rater Reconstruction Module

In order to further enhance the association between the *expertness* vector with the model prediction, and to capture the valuable inter-rater disagreement cues, a Multi-rater Reconstruction Module (MRM) is proposed to reconstruct the individual rater's annotation from the corresponding soft prediction P^1 and the given *expertness* vector V. Based on the reconstructed multi-rater's annotation, an uncertainty map that reflects the inter-rater variability is generated.

Specifically, as shown in Fig. 2(c), the initial prediction P^1 and the input image are concatenated and fed into an encoder-decoder network with VGG16 [41] as the feature encoder, since VGG architecture is well known for its superior capability that preserves the topological and perceptual features of the input image [22, 33]. The corresponding *expertness* vector V is applied at the bottleneck layer of the MRM via another EIM module. The decoder of MRM tries to reconstruct the annotations of individual raters via multiple 1×1 convolution layers (i.e.,$Conv_{1\times1}$) in the last layer.

Here, we employ a reconstruction loss, $loss_{\text{rec}}$, to measure the extent to which the reconstructed multi-raters' grading is similar to that of the real annotation marked by individual raters, which is defined as $loss_{\text{rec}} = \frac{1}{N}\sum_{i=1}^{N} L_{\text{BCE}}(S_i, \bar{S}_i)$. Here L_{BCE} denotes the binary cross entropy loss; N is the total number of experts; S_i and $\bar{S}_i \in \mathbb{R}^{W\times H\times C}$ denote the annotation marked by the i^{th} expert and the corresponding reconstructed prediction; W, H, and C denote the image width, height, and the number of channels, respectively.

To further improve the reconstruction performance of the MRM module, the fused soft label $GT^{\text{soft}} = \sum_i^N (V_i \cdot S_i)$, together with the given *expertness* vector, is also fed into the network to reconstruct the individual rater's grading. A consistency loss, $loss_{\text{con}}$, is proposed to enhance the coherence between the features extracted from the soft prediction P^1 and GT^{soft}, as $loss_{con} = \frac{1}{K}\sum_{i=1}^{K} \frac{1}{2}\left\|D_i - \bar{D}_i\right\|^2$. Here $\{\bar{D}_i\}_{i=1}^K$ and $\{D_i\}_{i=1}^K$ represent feature sets extracted from the encoder by using P^1 and GT^{soft} as input, respectively; K indicates the number of convolutional blocks where the features are extracted from and for the VGG16 backbone $K = 5$.

After reconstructing the individual rater's grading via the MRM, the uncertainty map of grading inconsistency can be estimated via the pixel-wise standard deviation of the multiple rater's predictions, using:

$$\mathrm{U}_{\text{map}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\bar{S}_i - \frac{1}{N}\sum_{i=1}^{N}\bar{S}_i\right)^2}. \quad (3)$$

The obtained uncertainty map is sent to the next module to further refine the initial coarse prediction P^1.

3.5. Multi-rater Perception Module

The grading inconsistency among multiple experts, i.e., the inter-rater variability, reflects the uncertainty or difficulty levels of different regions across the medical image. Thus, how to better take advantage of this information to further improve the segmentation performance is an important research problem. In this paper, we innovatively design a Multi-rater Perception Module (MPM), which can better capture and emphasize ambiguous regions by using the designed multi-branch soft attention mechanism.

Given the feature map F^1 obtained by the U-Net backbone and the estimated uncertainty map U_{map} obtained by the MRM, we use a spatial attention strategy [51] to emphasize the highly uncertain regions. However, the estimated uncertainty map might contain potential inaccuracy or incompleteness near the object boundaries, which may negatively affect the model performance if a 'hard' spatial attention is used. Therefore, we employ a 'soft' attention which aims to enlarge the coverage area of the uncertain regions, so as to effectively perceive and capture the disagreement cues among multiple raters. The soften operation can be formulated by:

$$Soft(U_{\text{map}}) = \Omega_{\max}(\mathcal{F}_{\text{Gauss}}(U_{\text{map}}, k), U_{\text{map}}), \quad (4)$$

where $\mathcal{F}_{\text{Gauss}}$ indicates a convolution operation with a Gaussian kernel k and zero bias, and $\Omega_{\max}$ indicates a maximum function to preserve the higher values between the Gaussian filtered map and the original uncertainty map U_{map}. In this paper, the size and standard deviation of the Gaussian kernel k are learnable through the model training procedure and initialized with 32 and 4, respectively.

Apart from the highly uncertain regions, the soft attention mechanism is applied on the initial coarse prediction map P^1 as well to enhance the highly certain regions for feature map F^1. In other words, both highly uncertain and certain regions are strengthened for F^1. For joint optic cup and disc segmentation task, F^1 is sent to four parallel branches with soft spatial attentions obtained from $\mathcal{A}_j = \{U_{\text{map}}^{\text{cup}}, U_{\text{map}}^{\text{disc}}, P_{\text{cup}}^1, P_{\text{disc}}^1\}_{j=1}^4$, as shown in Fig. 2(c). A skip connection is adopted between the original feature F^1 and the spatially enhanced features, so as to alleviate potential errors in the attention map being propagated to the network. The procedure is described as:

$$\tilde{F^j} = F^1 + Soft(\mathcal{A}_j) \otimes F^1, \quad (5)$$

where $\otimes$ denotes the pixel-wise multiplication operation and $\tilde{F^j}$ represents the refined feature from the j^{th} branch using the soft attention operation. The refined feature sets are further concatenated and fed to a $Conv_{1\times1}$ layer to ob-

tain the final segmentation prediction $\mathcal{M}$, as in:

$$\mathcal{M} = Conv_{1\times1}\left(Concat(\tilde{F^1}, \tilde{F^2}, \tilde{F^3}, \tilde{F^4})\right). \quad (6)$$

Finally, the total training loss $\mathcal{L}$ for the proposed MRNet framework is the combination of losses for the U-Net backbone, the MRM module and the MPM module, which can be represented as:

$$\begin{aligned}\mathcal{L} =& L_{\mathrm{BCE}}(P^1, GT^{\mathrm{soft}}) + L_{\mathrm{BCE}}(\mathcal{M}, GT^{\mathrm{soft}}) \\ &+ \alpha\, loss_{\mathrm{con}} + (1-\alpha)\, loss_{\mathrm{rec}}, \end{aligned} \quad (7)$$

where L_{BCE} denotes the binary cross entropy loss; α is a hyper-parameter that balances the weight of reconstruction loss $loss_{\mathrm{rec}}$ and consistency loss $loss_{\mathrm{con}}$ in the MRM module and empirically set as 0.7 in this work.

4. Experiments

4.1. Datasets

Extensive experiments are conducted to verify the effectiveness of the proposed framework on five different types of medical segmentation tasks with data from varied image modalities, including color fundus images, CT and MRI.

RIGA benchmark [1] is a publicly available dataset for retinal cup and disc segmentation, which contains in total of 750 color fundus images from three sources, including 460 images from MESSIDOR, 195 images from BinRushed and 95 images from Magrabia. Six glaucoma experts from different organizations labeled the optic cup and disc contour masks manually for the RIGA benchmark [1]. During model training, we select 195 samples from BinRushed and 460 samples from MESSIDOR as the training set, following [53]. The Magrabia set with 95 samples is selected as the test set to evaluate the model, which is not homologous to the training dataset. Parameters of the U-Net encoder are initialized with the model pre-trained on ImageNet [27].

QUBIQ benchmark [32], namely Quantification of Uncertainties in Biomedical Image Quantification Challenge, is a recently available challenge dataset specifically for the evaluation of inter-rater variability. QUBIQ contains four different segmentation datasets with CT and MRI modalities, including brain growth (one task, MRI, seven raters, 34 cases for training and 5 cases for testing), brain tumor (one task, MRI, three raters, 28 cases for training and 4 cases for testing), prostate (two subtasks, MRI, six raters, 33 cases for training and 15 cases for testing), kidney (one task, CT, three raters, 20 cases for training and 4 cases for testing).

4.2. Experimental Setup

4.2.1 Implementation Details

In our experiments, the main framework utilizes the U-Net architecture with ResNet34 as the backbone, and the MRM module utilizes the DeepLab-V3+ architecture with VGG-16 as the backbone. The network is implemented with the PyTorch platform and trained/tested on a Tesla P40 GPU with 24GB of memory. All training and test images are uniformly resized to the dimension of 256×256 pixels. The proposed network is trained in an end-to-end manner using the Adam optimizer [25], and it takes about 4 hours to train our model with a mini-batch size of 8 for 60 epochs. The initial learning rate is set to 1×10^{-4}.

4.2.2 Evaluation Metric

The target of the proposed network is to produce probability map $\mathcal{M}$ that can reflect the underlying inter-rater agreement/disagreement, i.e., calibrated predictions, for medical image segmentation. In order to better evaluate the calibrated model predictions, we use soft dice coefficient($\mathcal{D}$) / Intersection Over Union (IoU) metrics through multiple threshold levels, set as (0.1, 0.3, 0.5, 0.7, 0.9) in this paper, instead of using a single threshold (e.g., 0.5). At each threshold level, the predicted probability map $\mathcal{M}$ and soft GT GT^{soft} are binarized with the given threshold and then the $\mathcal{D}$ and IoU metrics are computed. The $\mathcal{D}$ and IoU scores obtained at multiple thresholds are averaged and then we obtain the soft metrics, denoted as $\mathcal{D}^{\mathrm{s}}$ and IoU^{s}, respectively. The higher the soft scores, the better calibrated the model performance.

4.3. Experimental Results

4.3.1 Performance of the Multi-rater Strategy

In order to verify that our multi-rater strategy can generate better calibrated segmentation maps under different given *expertness* conditions, we conduct quantitative experiments with different *expertness* setups on the RIGA test set in Table 2. Here, M1-M6 refer to the U-Net baseline model trained with the corresponding labels graded by Raters 1-6, respectively. In addition, three commonly used multi-rater strategies are employed to train the U-Net baseline model, including majority vote (i.e., U-Net baseline model [38] trained with the GT labels obtained by majority vote), label sampling [19] and multi-head strategies [16], denoted as MV-UNet, LS-UNet and MH-UNet, respectively, in Table 2. The performance of the comparison models is evaluated against various GT labels generated from different *expertness* vectors, including single rater condition (Raters1-6 raw gradings), random condition, average weight condition and STAPLE [50]. For the random condition, we randomly select three groups of results under different random *expertness* and report the average performance.

As listed in Table 2, the proposed MRNet consistently achieves superior performance under different conditions, reflecting the dynamic representation capability of the MRNet by incorporating the expertise cues of individual raters. In addition, it is worth noting that our approach achieves the

Table 2. Quantitative results with different strategies on the RIGA test set under various expertise levels and ground-truths. The GTs are set as individual rater mode (Rater1-6), fused using random conditions, majority vote of average weight and STAPLE strategy [50]. Here, we use soft metrics ($\mathcal{D}^s_{disc}$ (%), $\mathcal{D}^s_{cup}$ (%)) to evaluate these results, where the best three results are shown in **bold**, red and blue, respectively.

Final Label	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Random	Average	STAPLE
Expertness	[1,0,0,0,0,0]	[0,1,0,0,0,0]	[0,0,1,0,0,0]	[0,0,0,1,0,0]	[0,0,0,0,1,0]	[0,0,0,0,0,1]	[-,-,-,-,-,-]	[1,1,1,1,1,1]	[1,1,1,1,1,1]
M1 (Rater1)	(95.11, 78.96)	(93.88, 76.68)	(95.24, 77.52)	(95.15, 75.75)	(95.60, 77.83)	(95.55, 74.13)	(96.94, 82.16)	(97.10, 83.48)	(96.01, 83.43)
M2 (Rater2)	(**95.74**, 78.82)	(**95.48**, 80.65)	(95.38, 77.12)	(95.12, 77.42)	(95.01, 78.00)	(95.27, 73.80)	(96.85, 82.41)	(96.77, 83.10)	(95.80, 82.96)
M3 (Rater3)	(95.30, 77.02)	(94.63, 77.31)	(**96.21**, **82.49**)	(94.73, 76.14)	(94.14, 76.40)	(95.09, 74.85)	(96.57, 81.24)	(96.66, 82.04)	(95.49, 80.97)
M4 (Rater4)	(95.20, 76.47)	(94.38, 80.42)	(94.81, 76.69)	(**96.58**, **86.88**)	(95.52, 72.31)	(95.39, 68.95)	(96.99, 77.45)	(97.01, 78.68)	(96.12, 85.49)
M5 (Rater5)	(95.18, 78.37)	(94.82, 76.73)	(95.05, 78.13)	(95.18, 72.67)	(95.34, **80.53**)	(95.97, 74.44)	(96.60, 79.13)	(96.68, 79.58)	(95.64, 75.22)
M6 (Rater6)	(95.05, 77.72)	(94.64, 75.35)	(95.39, 75.10)	(95.16, 69.90)	(95.09, 78.31)	(**96.34**, **78.60**)	(97.00, 79.42)	(96.99, 79.01)	(95.77, 72.73)
MV-UNet [38]	(94.87, 78.68)	(95.47, 77.62)	(95.12, 76.67)	(94.82, 76.75)	(95.44, 77.76)	(95.71, 78.54)	(97.11, 82.42)	(97.03, 82.88)	(95.94, 84.22)
LS-UNet[19]	(94.85, 76.92)	(94.26, 76.03)	(94.89, 75.73)	(95.20, 77.77)	(95.10, 74.02)	(95.13, 71.02)	(96.62, 80.95)	(96.90, 82.41)	(94.99, 81.24)
MH-UNet [16]	(94.71, 81.25)	(94.73, 80.27)	(95.77, 78.97)	(95.71, 83.89)	(95.52, 78.91)	(96.11, 76.78)	(96.37, 83.31)	(96.81, 82.17)	(96.15, 81.52)
Ours	(95.35, **81.77**)	(94.81, **81.18**)	(95.80, 79.23)	(95.96, 84.46)	(**95.90**, 79.04)	(95.76, 76.20)	(**97.28**, **85.65**)	(**97.55**, **87.20**)	(**96.26**, **86.37**)

best performance under majority vote (i.e., average weight condition), with a large performance margin over all the other models, including the MV-UNet which is specifically trained with the majority vote consensus labels. These empirical experiments demonstrate the effectiveness of the proposed framework which is tailored for medical image segmentations with multi-rater annotations, by taking advantage of the proposed dynamic *expertness* inferring and multi-rater agreement modeling.

4.3.2 Comparisons with State-of-the-arts

To demonstrate the advantage of the proposed MRNet, we compare our method with the state-of-the-art (SOTA) methods for joint optic cup and disc segmentation task. We use the publicly released code with default parameters to retrain the SOTA methods, with the same training/test set as that of ours for a fair comparison.

Table 3 quantitatively compares our framework with five SOTA cup/disc segmentation methods, including ResUnet [53], CENet [15], AGNet [58], BEAL [45] and pOSAL [46] on the RIGA test set. As shown in Table 3, our proposed MRNet consistently achieves superior performance compared with SOTA optic cup/disc segmentation methods. The performance improvement is especially prominent for the retinal cup segmentation where the inter-observer variability is more significant, with an increase of 1.2% for soft dice coefficient value over the current best method.

Fig. 4 shows two typical examples generated by our MRNet and other SOTA methods. It is obvious that the probability map generated by the proposed model is better calibrated compared with other methods, especially for the ambiguous regions among different experts. Thus, the predictions generated by the proposed MRNet is able to better reflect the underlying dis-/agreement among multiple experts.

4.3.3 Ablation Studies

In this section, ablation studies are performed over each component of the proposed MRNet, including the EIM, MRM and MPM, as listed in Table 4 and Table 5. All experiments are evaluated using the soft GT obtained with

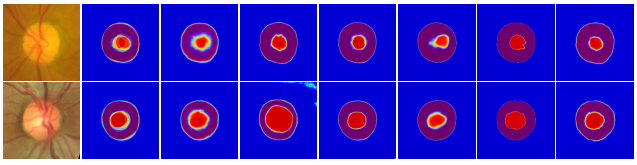

Figure 4. Visual comparisons of our MRNet with the state-of-the-arts for joint optic cup and disc segmentation tasks.

Table 3. Quantitative comparisons with the state-of-the-art methods for optic cup and disc segmentation on the Magrabia dataset.

	$\mathcal{D}^s_{\text{disc}}$ (%)	$\mathcal{D}^s_{\text{cup}}$ (%)	IoU^s_{disc} (%)	IoU^s_{cup} (%)
AGNet [58]	96.31	72.05	92.93	59.44
CENet [15]	96.55	81.82	93.38	71.03
ResUnet [53]	96.75	85.38	93.75	75.76
pOSAL [46]	95.85	84.07	92.12	74.40
BEAL [45]	97.08	85.97	94.38	77.18
MRNet (ours)	**97.55**	**87.20**	**95.24**	**78.62**

majority vote, i.e., the average weight *expertness* condition. In Table 4, as we sequentially adding the proposed modules on top of the U-Net baseline, the model performance is gradually improved, especially for that of the optic cup. Firstly, by integrating the EIM with ConvLSTM into the UNet baseline, the $\mathcal{D}^s_{\text{cup}}$ value is increased by 1.0%. Compared to the direct condition operation by concatenating *expertness* with feature maps, the EIM with ConvLSTM operation achieves better performance (Table 4(b) vs. (c)). This indicates that the introduction of multi-rater expertise knowledge via EIM with ConvLSTM improves the dynamic representation capability of the model and the exploitation of multi-rater annotations can arrive at better calibrated predictions. Additionally, in order to effectively utilize the multi-rater cues for calibrating the segmentation results, the MRM and MPM modules are specifically designed to reconstruct the raw multi-rater gradings and further to utilize the multi-rater (dis-)agreement cues, which boosts the $\mathcal{D}^s_{\text{cup}}$ metric by 2.0% and 1.5%, respectively.

To further investigate the influence of individual losses and operations in the MRM and MPM, i.e., the MAM module, a set of ablation studies is conducted for the reconstruction loss ($loss_{\text{rec}}$), consistency loss ($loss_{\text{con}}$) and soft atten-

tion operation ($Soft_{\text{op}}$). In Table 5, the reconstruction loss significantly improves the $\mathcal{D}^{\text{s}}_{\text{cup}}$ by 1.2%, reflecting the necessity of reconstructing raw multi-rater's grading from the fused soft GT/prediction. By adding the consistency loss to constrain the features extracted from soft GT and coarse prediction, the $\mathcal{D}^{\text{s}}_{\text{cup}}$ is further improved by 0.8%. Moreover, comparing Table 5 (iv) and (v), the soft attention operation further boosts the $\mathcal{D}^{\text{s}}_{\text{cup}}$ by 1.2% compared with using 'hard' attention, achieving the final $\mathcal{D}^{\text{s}}_{\text{cup}}$ score of 87.2%. This verifies that the proposed soft attention mechanism can better emphasize both certain and uncertain regions and further to improve the calibration performance of the model. We also investigate the influence of different U-Net backbones and the α hyper-parameter. With a stronger backbone (ResNet101 vs. ResNet34), the model performance can be further improved (88.45% vs 87.20%), in terms of $\mathcal{D}^{\text{s}}_{\text{cup}}(\%)$. When α is set as 0.3, 0.7, 0.9, the corresponding $\mathcal{D}^{\text{s}}_{\text{cup}}(\%)$ is 86.17%, 87.20% and 86.87%, respectively.

Table 4. Ablation analysis on the RIGA test set.

Module						Average *Expertness*	
Index	Baseline	EIM	ConvLSTM	MRM	MPM	$\mathcal{D}^{\text{s}}_{\text{disc}}$ (%)	$\mathcal{D}^{\text{s}}_{\text{cup}}$ (%)
(a)	✓					97.03	82.88
(b)	✓	✓	×			97.07	83.19
(c)	✓	✓	✓			97.16	83.74
(d)	✓	✓	✓	✓		97.52	85.75
(e)	✓	✓	✓	✓	✓	97.55	87.20

Table 5. Ablation analysis of our MAM on the RIGA test set. Here, all experiments are based on UNet baseline + EIM.

MAM						Average *Expertness*	
No.	Table 4 (b)	$loss_{\text{rec}}$	$loss_{\text{con}}$	MPM	$Soft_{\text{op}}$	$\mathcal{D}^{\text{s}}_{\text{disc}}$ (%)	$\mathcal{D}^{\text{s}}_{\text{cup}}$ (%)
(i)	✓					97.16	83.74
(ii)	✓	✓				97.39	84.94
(iii)	✓	✓	✓			97.52	85.75
(iv)	✓	✓	✓	✓	×	97.54	86.05
(v)	✓	✓	✓	✓	✓	97.55	87.20

4.3.4 Generalization Capability

To further verify the effectiveness and generalization capability of the proposed MRNet, a generalization experiment is conducted on a recently released QUBIQ dataset for four types of medical image segmentation tasks that contain both CT and MRI modalities. Several commonly used multi-rater strategies are adopted for comparison, including U-Net [38] based on majority vote (MV-UNet), label sampling [19] (LS-UNet) and multiple head strategies [16] (MH-UNet). As listed in Table 6, the proposed MRNet achieves better calibrated performance compared with other commonly used methods and multi-rater strategies. These quantitative results again verify that the underlying agreement/disagreement information among multiple experts regarding the pathological region are beneficial to improve calibrated segmentation accuracy through our multi-rater agreement modeling. Several representative examples of the comparison methods for four different types of medical image segmentation are visualized in Fig. 5.

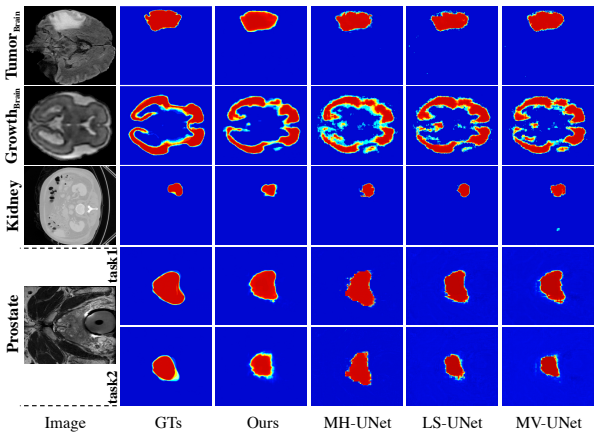

Figure 5. Segmentation results of different strategies for four different medical segmentation tasks on the QUBIQ dataset.

Table 6. Quantitative evaluation of five medical segmentation subtasks with multi-rater modeling on the QUBIQ dataset, including the segmentation of kidney ($\mathcal{D}^{\text{s}}_{\text{kidney}}$), brain growth ($\mathcal{D}^{\text{s}}_{\text{brain}}$), brain tumor ($\mathcal{D}^{\text{s}}_{\text{tumor}}$) and two prostate tasks ($\mathcal{D}^{\text{s}}_{\text{pros1}}$ and $\mathcal{D}^{\text{s}}_{\text{pros2}}$).

(%)	$\mathcal{D}^{\text{s}}_{\text{kidney}}$	$\mathcal{D}^{\text{s}}_{\text{brain}}$	$\mathcal{D}^{\text{s}}_{\text{tumor}}$	$\mathcal{D}^{\text{s}}_{\text{pros1}}$	$\mathcal{D}^{\text{s}}_{\text{pros2}}$
FCN [31]	70.03	80.99	83.12	84.55	67.81
MC Dropout [14]	72.93	82.91	86.17	86.40	70.95
FPM [61]	72.17	-	-	-	-
DAF [47]	-	-	-	85.98	72.87
MV-UNet [38]	70.65	81.77	84.03	85.18	68.39
LS-UNet [19]	72.31	82.79	85.85	86.23	69.05
MH-UNet [16]	73.44	83.54	86.74	87.03	75.61
MRNet (ours)	**74.97**	**84.31**	**88.40**	**87.27**	**76.01**

5. Conclusion

In this work, we focus on the utilization of rich annotation information from multiple experts, which are relatively less-explored but widely presented in the medical image grading procedure. We proposed to incorporate the multi-rater (dis-)agreement cues in our MRNet framework and generate calibrated model predictions that better reflected the underlying agreement among multiple experts. This was achieved by the use of an expertise-aware inferring module to explicitly integrate graders expertise cues into high-level semantic features, as well as a multi-rater agreement modeling module to reconstruct gradings of individual raters and refine the coarse prediction to form the final calibrated segmentation maps. Extensive empirical experiments demonstrated the overall superior performance of our MRNet on a range of medical image segmentation tasks over diverse image modalities.

Acknowledgement. This work was funded by the Key-Area Research and Development Program of Guangdong Province, China (No. 2018B010111001), National Key R&D Program of China (2018YFC2000702) and the Scientific and Technical Innovation 2030-'New Generation Artificial Intelligence' Project (No. 2020AAA0104100), University of Alberta Start-up Grant, and NSERC Discovery Grants (No. RGPIN-2019-04575).

References

[1] Ahmed Almazroa, Sami Alodhayb, Essameldin Osman, Eslam Ramadan, Mohammed Hummadi, Mohammed Dlaim, Muhannad Alkatee, Kaamran Raahemifar, and Vasudevan Lakshminarayanan. Agreement among ophthalmologists in marking the optic disc and optic cup in fundus images. *International Ophthalmology*, 37(3):701–717, 2017. 3, 6

[2] Christian F Baumgartner, Kerem C Tezcan, Krishna Chaitanya, Andreas M Hötker, Urs J Muehlematter, Khoschy Schawkat, Anton S Becker, Olivio Donati, and Ender Konukoglu. PHiSeg: Capturing uncertainty in medical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 119–127. Springer, 2019. 2, 3

[3] Anton S Becker, Krishna Chaitanya, Khoschy Schawkat, Urs J Muehlematter, Andreas M Hötker, Ender Konukoglu, and Olivio F Donati. Variability of manual segmentation of the prostate in axial T2-weighted MRI: A multi-reader study. *European Journal of Radiology*, 121:108716, 2019. 1

[4] Cheng Chen, Qi Dou, Yueming Jin, Hao Chen, Jing Qin, and Pheng-Ann Heng. Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 447–456. Springer, 2019. 2

[5] Geng Chen, Dehui Xiang, Bin Zhang, Haihong Tian, Xiaoling Yang, Fei Shi, Weifang Zhu, Bei Tian, and Xinjian Chen. Automatic pathological lung segmentation in low-dose CT image using eigenspace sparse shape composition. *IEEE Transactions on Medical Imaging*, 38(7):1736–1749, 2019. 2

[6] Shengcong Chen, Changxing Ding, and Minfeng Liu. Dual-force convolutional neural networks for accurate brain tumor segmentation. *Pattern Recognition*, 88:90–100, 2019. 2

[7] Wenting Chen, Shuang Yu, Junde Wu, Kai Ma, Cheng Bian, Chunyan Chu, Linlin Shen, and Yefeng Zheng. TR-GAN: Topology ranking gan with triplet loss for retinal artery/vein classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 616–625. Springer, 2020. 1

[8] Jun Cheng, Jiang Liu, Yanwu Xu, Fengshou Yin, Damon Wing Kee Wong, Ngan-Meng Tan, Dacheng Tao, Ching-Yu Cheng, Tin Aung, and Tien Yin Wong. Superpixel classification based optic disc and optic cup segmentation for glaucoma screening. *IEEE Transactions on Medical Imaging*, 32(6):1019–1032, 2013. 2

[9] Li Cheng and Terry Caelli. Unsupervised image segmentation: A Bayesian approach. In *Proceedings of 16th International Conference on Vision Interface, Halifax, Canada*, pages 251–257, 2003. 2

[10] Li Cheng, Ning Ye, Weimiao Yu, and Andre Cheah. Discriminative segmentation of microscopic cellular images. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 637–644. Springer, 2011. 2

[11] Jason J Corso, Eitan Sharon, Shishir Dube, Suzie El-Saden, Usha Sinha, and Alan Yuille. Efficient multilevel brain tumor segmentation with integrated Bayesian model classification. *IEEE Transactions on Medical Imaging*, 27(5):629–640, 2008. 1

[12] Huazhu Fu, Jun Cheng, Yanwu Xu, Damon Wing Kee Wong, Jiang Liu, and Xiaochun Cao. Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. *IEEE Transactions on Medical Imaging*, 37(7):1597–1605, 2018. 2

[13] Huazhu Fu, Fei Li, Yanwu Xu, Jingan Liao, Jian Xiong, Jianbing Shen, Jiang Liu, and Xiulan Zhang. A retrospective comparison of deep learning to manual annotations for optic disc and optic cup segmentation in fundus photos. *medRxiv*, 2020. 1, 3

[14] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059, 2016. 8

[15] Zaiwang Gu, Jun Cheng, Huazhu Fu, Kang Zhou, Huaying Hao, Yitian Zhao, Tianyang Zhang, Shenghua Gao, and Jiang Liu. CE-Net: Context encoder network for 2D medical image segmentation. *IEEE Transactions on Medical Imaging*, 38(10):2281–2292, 2019. 2, 7

[16] Melody Guan, Varun Gulshan, Andrew Dai, and Geoff Hinton. Who said what: Modeling individual labelers improves classification. In *AAAI Conference on Artificial Intelligence*, pages 3109–3118, 2018. 2, 3, 6, 7, 8

[17] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 35:18–31, 2017. 1

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645, 2016. 4

[19] Martin Holm Jensen, Dan Richter Jørgensen, Raluca Jalaboi, Mads Eiler Hansen, and Martin Aastrup Olsen. Improving uncertainty estimation in convolutional neural networks using inter-rater agreement. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 540–548, 2019. 2, 3, 6, 7, 8

[20] Wei Ji, Jingjing Li, Miao Zhang, Yongri Piao, and Huchuan Lu. Accurate rgb-d salient object detection via collaborative learning. *European Conference on Computer Vision*, 2020. 1

[21] Haozhe Jia, Yong Xia, Yang Song, Donghao Zhang, Heng Huang, Yanning Zhang, and Weidong Cai. 3D APA-Net: 3D adversarial pyramid anisotropic convolutional network for prostate segmentation in MR images. *IEEE Transactions on Medical Imaging*, 39(2):447–457, 2019. 2

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016. 5

[23] Leo Joskowicz, D Cohen, N Caplan, and J Sosna. Inter-observer variability of manual contour delineation of structures in CT. *European Radiology*, 29(3):1391–1399, 2019. 1, 3

[24] Alain Jungo, Raphael Meier, Ekin Ermis, Marcela Blatti-Moreno, Evelyn Herrmann, Roland Wiest, and Mauricio Reyes. On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 682–690. Springer, 2018. 2, 3

[25] Diederik P Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 6

[26] Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey de Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic U-Net for segmentation of ambiguous images. In *Advances in Neural Information Processing Systems*, pages 6965–6975, 2018. 2

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012. 6

[28] Sajeesh Kumar, Antonio Giubilato, William Morgan, Ludmila Jitskaia, Chris Barry, Max Bulsara, Ian J Constable, and Kanagasingam Yogesan. Glaucoma screening: analysis of conventional and telemedicine-friendly devices. *Clinical & Experimental Ophthalmology*, 35(3):237–243, 2007. 1, 3

[29] Ge Li, Changsheng Li, Chan Zeng, Peng Gao, and Guotong Xie. Region focus network for joint optic disc and cup segmentation. *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, 34(1):751–758, 2020. 2

[30] Quande Liu, Qi Dou, Lequan Yu, and Pheng Ann Heng. MS-Net: Multi-site network for improving prostate segmentation with heterogeneous MRI data. *IEEE Transactions on Medical Imaging*, 2020. 1, 2

[31] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 8

[32] Bjoern Menze, Leo Joskowicz, Spyridon Bakas, Andras Jakab, Ender Konukoglu, Anton Becker, and et al. https://qubiq.grand-challenge.org. In *Quantification of Uncertainties in Biomedical Image Quantification Challenge at MICCAI*, 2020. 6

[33] Agata Mosinska, Pablo Marquez-Neila, Mateusz Koziński, and Pascal Fua. Beyond the pixel-wise loss for topology-aware delineation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3136–3145, 2018. 5

[34] Andriy Myronenko. 3D MRI brain tumor segmentation using autoencoder regularization. In *International MICCAI Brainlesion Workshop*, pages 311–320. Springer, 2018. 1, 2

[35] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7254–7263, 2019. 1

[36] Harry A Quigley and Aimee T Broman. The number of people with glaucoma worldwide in 2010 and 2020. *British Journal of Ophthalmology*, 90(3):262–267, 2006. 1

[37] Maryanne Romero, Vivien Lim, and Seng Chee Loon. Reliability of graders and comparison with an automated algorithm for vertical cup-disc ratio grading in fundus photographs. *Ann Acad Med Singapore*, 48:282–9, 2019. 1

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 234–241, 2015. 3, 4, 6, 7, 8

[39] Mike Schaekermann, Graeme Beaton, Minahz Habib, Andrew Lim, Kate Larson, and Edith Law. Understanding expert disagreement in medical data analysis through structured adjudication. *Proceedings of the ACM on Human-Computer Interaction*, 3:1–23, 2019. 1, 3

[40] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai kin Wong, and Wang chun Woo. Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems*, pages 802–810, 2015. 4

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2015. 5

[42] Carole H Sudre, Beatriz Gomez Anson, Silvia Ingala, Chris D Lane, Daniel Jimenez, Lukas Haider, Thomas Varsavsky, Ryutaro Tanno, Lorna Smith, Sébastien Ourselin, et al. Let's agree to disagree: Learning highly debatable multirater labelling. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 665–673. Springer, 2019. 2

[43] Yih Chung Tham, Xiang Li, Tien Y Wong, Harry A Quigley, Tin Aung, and Ching Yu Cheng. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology*, 121(11):2081–2090, 2014. 1

[44] Zhuowen Tu and Xiang Bai. Auto-context and its application to high-level vision tasks and 3D brain image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(10):1744–1757, 2009. 1

[45] Shujun Wang, Lequan Yu, Kang Li, Xin Yang, Chi-Wing Fu, and Pheng-Ann Heng. Boundary and entropy-driven adversarial learning for fundus image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 102–110, 2019. 2, 7

[46] Shujun Wang, Lequan Yu, Xin Yang, Chi-Wing Fu, and Pheng-Ann Heng. Patch-based output space adversarial learning for joint optic disc and cup segmentation. *IEEE Transactions on Medical Imaging*, 38(11):2485–2495, 2019. 7

[47] Yi Wang, Zijun Deng, Xiaowei Hu, Lei Zhu, Xin Yang, Xuemiao Xu, Pheng-Ann Heng, and Dong Ni. Deep attentional features for prostate segmentation in ultrasound. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 523–530. Springer, 2018. 8

[48] Yi Wang, Haoran Dou, Xiaowei Hu, Lei Zhu, Xin Yang, Ming Xu, Jing Qin, Pheng-Ann Heng, Tianfu Wang, and Dong Ni. Deep attentive features for prostate segmentation in 3D transrectal ultrasound. *IEEE Transactions on Medical Imaging*, 38(12):2768–2778, 2019. 2

[49] Yan Wang, Yuyin Zhou, Wei Shen, Seyoun Park, Elliot K Fishman, and Alan L Yuille. Abdominal multi-organ segmentation with organ-attention networks and statistical fusion. *Medical Image Analysis*, 55:88–102, 2019. 1

[50] Simon K Warfield, Kelly H Zou, and William M Wells. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 2004. 2, 6, 7

[51] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*, pages 3–19, 2018. 5

[52] Lequan Yu, Xin Yang, Hao Chen, Jing Qin, Pheng-Ann Heng, et al. Volumetric ConvNets with mixed residual connections for automated prostate segmentation from 3D MR images. In *AAAI Conference on Artificial Intelligence*, volume 17, pages 36–72, 2017. 1

[53] Shuang Yu, Di Xiao, Shaun Frost, and Yogesan Kanagasingam. Robust optic disc and cup segmentation with deep learning for glaucoma detection. *Computerized Medical Imaging and Graphics*, 74:61–71, 2019. 6, 7

[54] Shuang Yu, Hong-Yu Zhou, Kai Ma, Cheng Bian, Chunyan Chu, Hanruo Liu, and Yefeng Zheng. Difficulty-aware glaucoma classification with multi-rater consensus modeling. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 741–750. Springer, 2020. 2, 3

[55] Miao Zhang, Wei Ji, Yongri Piao, Jingjing Li, Yu Zhang, Shuang Xu, and Huchuan Lu. LFNet: Light field fusion network for salient object detection. *IEEE Transactions on Image Processing*, 29:6276–6287, 2020. 1

[56] Miao Zhang, Jingjing Li, Wei Ji, Yongri Piao, and Huchuan Lu. Memory-oriented decoder for light field salient object detection. In *Advances in Neural Information Processing Systems*, pages 896–906, 2019. 1

[57] Qinming Zhang, Luyan Liu, Kai Ma, Cheng Zhuo, and Yefeng Zheng. Cross-denoising network against corrupted labels in medical image segmentation with domain shift. In *International Joint Conference on Artificial Intelligence*, 2020. 2

[58] Shihao Zhang, Huazhu Fu, Yuguang Yan, Yubing Zhang, Qingyao Wu, Ming Yang, Mingkui Tan, and Yanwu Xu. Attention guided network for retinal image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 797–805, 2019. 2, 7

[59] He Zhao, Huiqi Li, and Li Cheng. Improving retinal vessel segmentation with joint local loss by matting. *Pattern Recognition*, 98:107068, 2020. 1, 2

[60] He Zhao, Huiqi Li, Sebastian Maurer-Stroh, Yuhong Guo, Qiuju Deng, and Li Cheng. Supervised segmentation of unannotated retinal fundus images by synthesis. *IEEE Transactions on Medical Imaging*, 38(1):46–56, 2018. 2

[61] Yuyin Zhou, Lingxi Xie, Wei Shen, Yan Wang, Elliot K Fishman, and Alan L Yuille. A fixed-point model for pancreas segmentation in abdominal CT scans. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 693–701. Springer, 2017. 8