Paper List

CVPR2021	Learning Calibrated Medical Image Segmentation via Multi-rater Agreement Modeling
MICCAI2020	QUBIQ Challenge

Learning Calibrated Medical Image Segmentation via Multi-rater Agreement Modeling

Information

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

Background

- Accurate segmentation is important
- Medical images are often independently annotated by a group of experts or raters
- Inter-observer variability often leads to challenges in segmenting highly uncertain regions
- Common Label Fusion Strategies, e.g. majority vote & STAPLE, over-confident
- Necessary for automated systems to consider a proper segmentation strategy that reflects the underlying (dis-)agreement among multiple experts

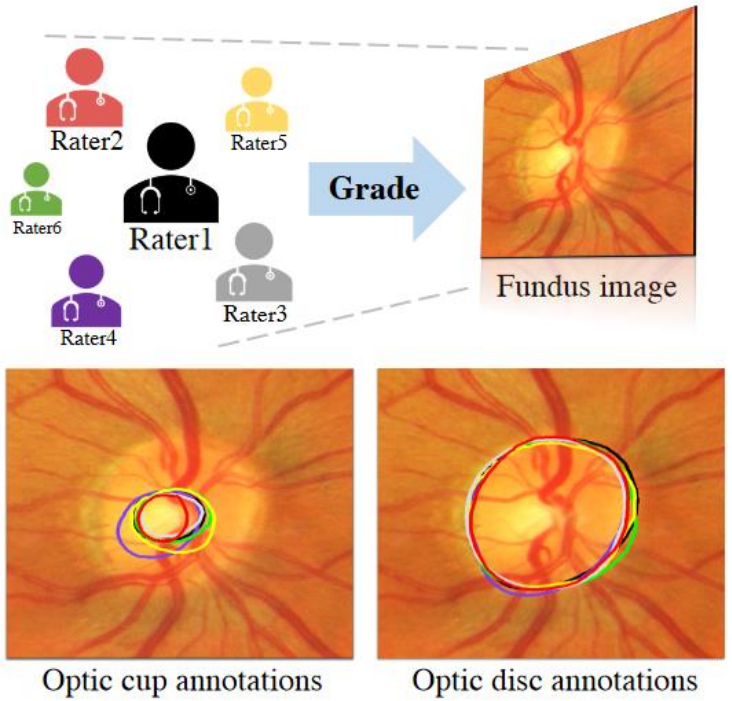


Figure 1. **Top:** an exemplar medical image grading scenario conducted by multiple raters with different expertise levels. **Bottom:** visualization of optic cup and disc annotations of the above raters.

Previous Work

1. Medical Image Seg:
   - Train by retaining unique GT labels.
   - The valuable multi-rater labels with inter-rater variability are not well-exploited.
2. Multi-rater Strategies：
   - Label Sampling: Sample labels randomly from the multi-rater labeling pool during each training iteration [1, 2, 3]
   - Predict the gradings of each rater individually and learn the corresponding weights for final prediction [4]
   - A multi-branch structure to generate three predictions under different sensitivity settings, to leverage multi-rater consensus information for glaucoma classification [5]

Motivation

- Preliminary Experiment
  - ❖Target: Quantitatively demonstrate the difference in experts' grading preferences and levels of expertise
  - ❖Task: Optical cup segmentation
  - ❖Model: U-Net
  - ❖Dataset: RIGA benchmark dataset[6]
  - ❖Result: Table 1
- Findings: (consistent with analysis in [6])
  1) Individual expert has specific and consistent grading patterns;
  2) Expertise levels among a group of graders are usually different from one to the other.

Motivation

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

Contributions

1. An expertise-aware inferring module (EIM) is devised to embed the expertise level of individual raters as prior knowledge, to form high-level semantic features
2. Capable of reconstructing multi-rater gradings from coarse predictions, with the multi-rater (dis-)agreement cues being further exploited to improve the segmentation performance

Novality of MRNet

1. First in producing calibrated predictions under different expertise levels for medical image segmentation
2. Real-time (29 frame per second) at inference stage, making it practically appealing for many real-world applications

Architecture of MRNet

1. EIM (Expertise-aware Inferring Module) Prior knowledge

2. MAM (Multi-rater Agreement Modeling) consists of 2parts:
   - MRM (Multi-rater Reconstruction Module): Reconstruct the raw multi-rater gradings from prior and the soft prediction
   - MPM (Multi-rater Perception Module): Better utilize the rich cues among multi-rater (dis-)agreements

Architecture of MRNet

Stage 1: Coarse prediction

- Input: fundus image & expertness vector
- Output: coarse prediction P^1 & Feature F_1
- Model: U-Net(Encoder part: Pretrained ResNet34) + EIM

Stage 2: MAM

- Input: expertness vector & P^1 & F_1
- Output: Final refined prediction M
- Model:
  - MRM ----- VGG16
  - MPM ----- SoftAttention

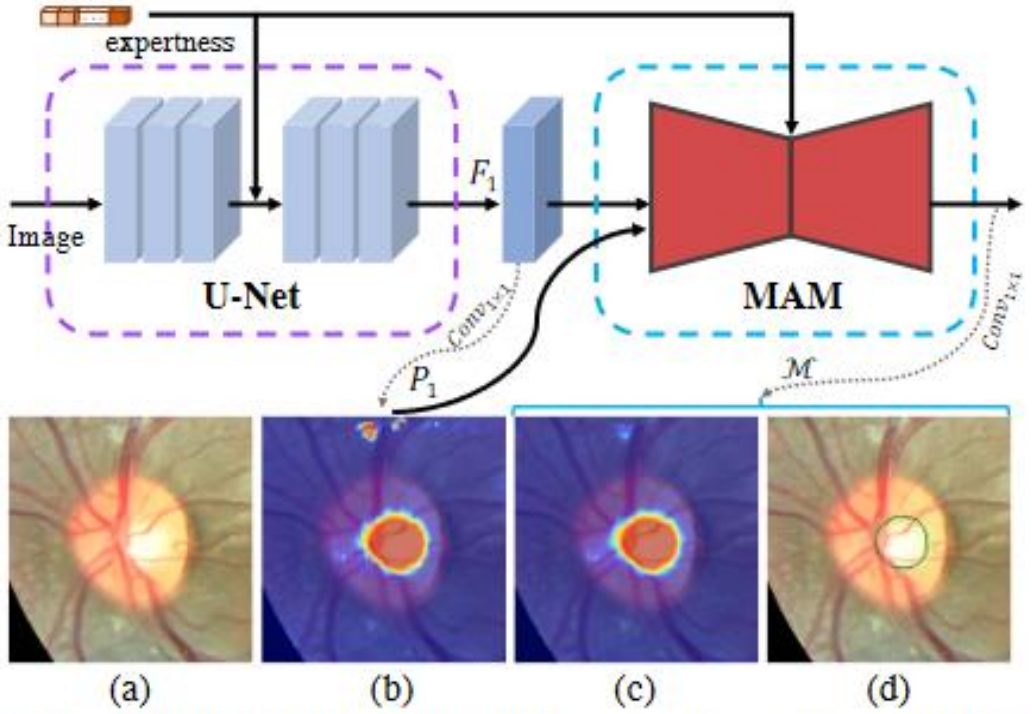


Figure 3. Intermediate visual results in the processing pipeline of our MRNet framework. (a) Input fundus image. (b) Heat map of the initial cup prediction P^1. (c) Heat map of the final refined cup prediction $\mathcal{M}$. (d) Segmentation boundary of the cup prediction (green) and ground-truth (black).

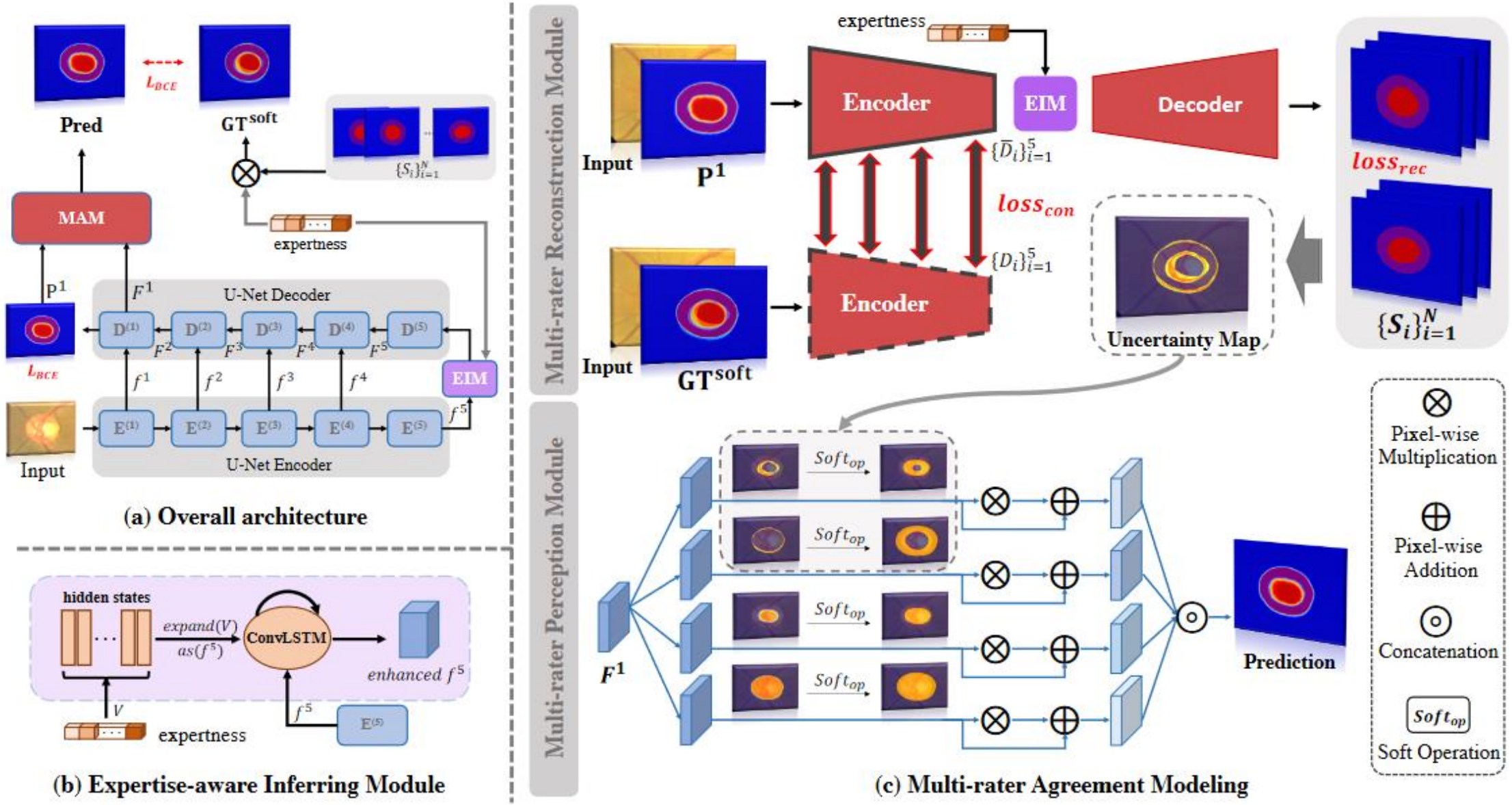
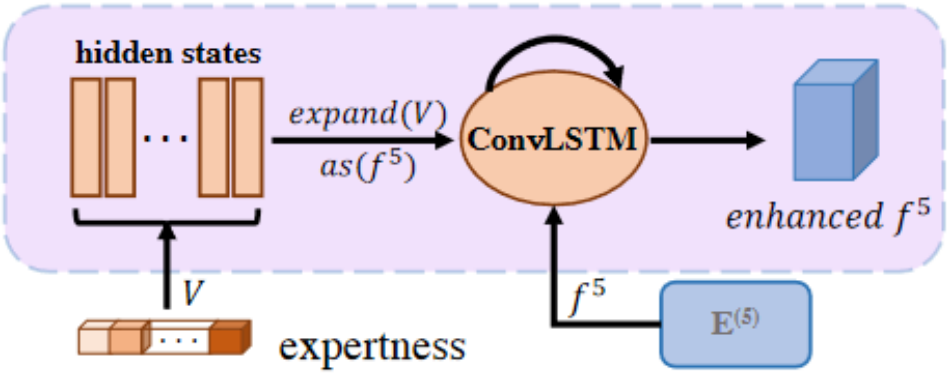


Figure 2. An illustration of our MRNet framework, which starts from (a) an overview of the processing pipeline, and continues with zoomed-in diagrams of individual modules, including (b) the Expertise-aware Inferring Module (EIM), and (c) the Multi-rater Agreement Modeling (MAM) that consists of the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

Architecture

EIM

- V: A normalized expertness vector
  - 3 modes:
    1. the majority vote mode
    2. single rater mode
    3. random weight assignment
  - Advantages:
    1. associate the influence of individual raters on the final soft predictio
    2. effective data augmentation strategy



(b) Expertise-aware Inferring Module

- ConvLSTM: $h_0 = V$

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

Architecture

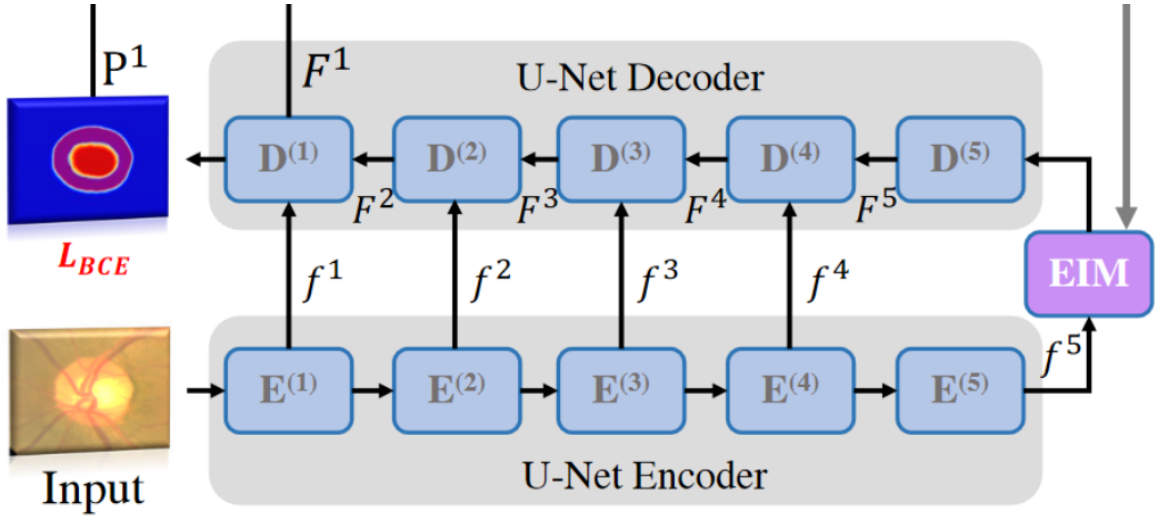
Stage 1: Coarse prediction

- Input:

  Fundus Images x & Expertness Vector V
- Output:

  Coarse Prediction P^1 & Feature F^1
- Model:

  U-Net(Encoder part: Pretrained ResNet34) + EIM



- Loss: BCE Loss (P^1,GT^{soft}) + consistency loss($loss_{con}$)

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

Architecture

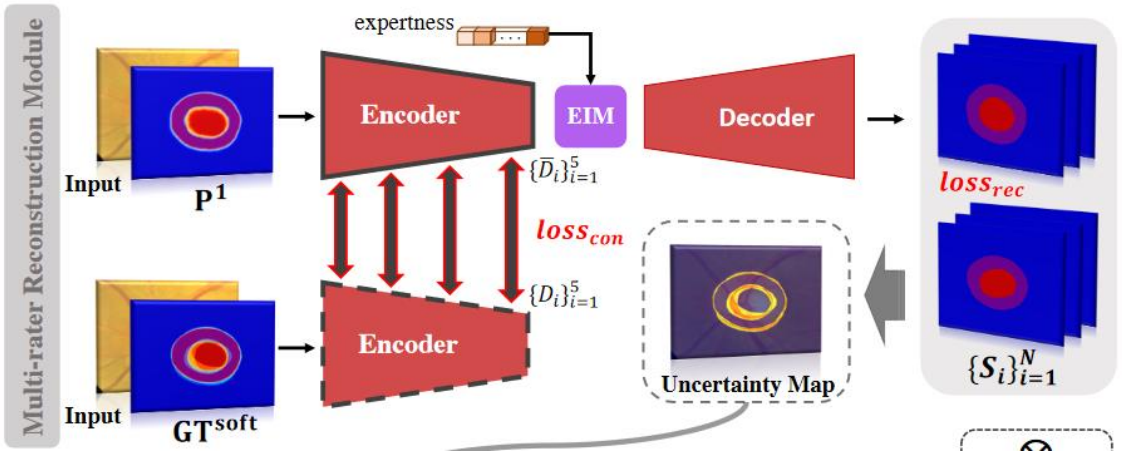
MRM

- Model: VGG16 as encoder + EIM + ASPP Decoder
- Input: concat(x, P^1)
- Output: $\{\bar{S}_i\}^N$,N =# of Raters & Uncertainty Map
- Loss: :Produce 2 losses
  - $loss_{rec}$ for MRM

$$\frac{1}{N}\sum_{i=1}^{N} L_{\mathrm{BCE}}(S_i, \bar{S}_i)$$

  - $loss_{con}$ for U-Net in Step 1, K = # of Conv blocks in Encoder

$$\frac{1}{K}\sum_{i=1}^{K} \frac{1}{2}\left\|D_i - \bar{D}_i\right\|^2$$

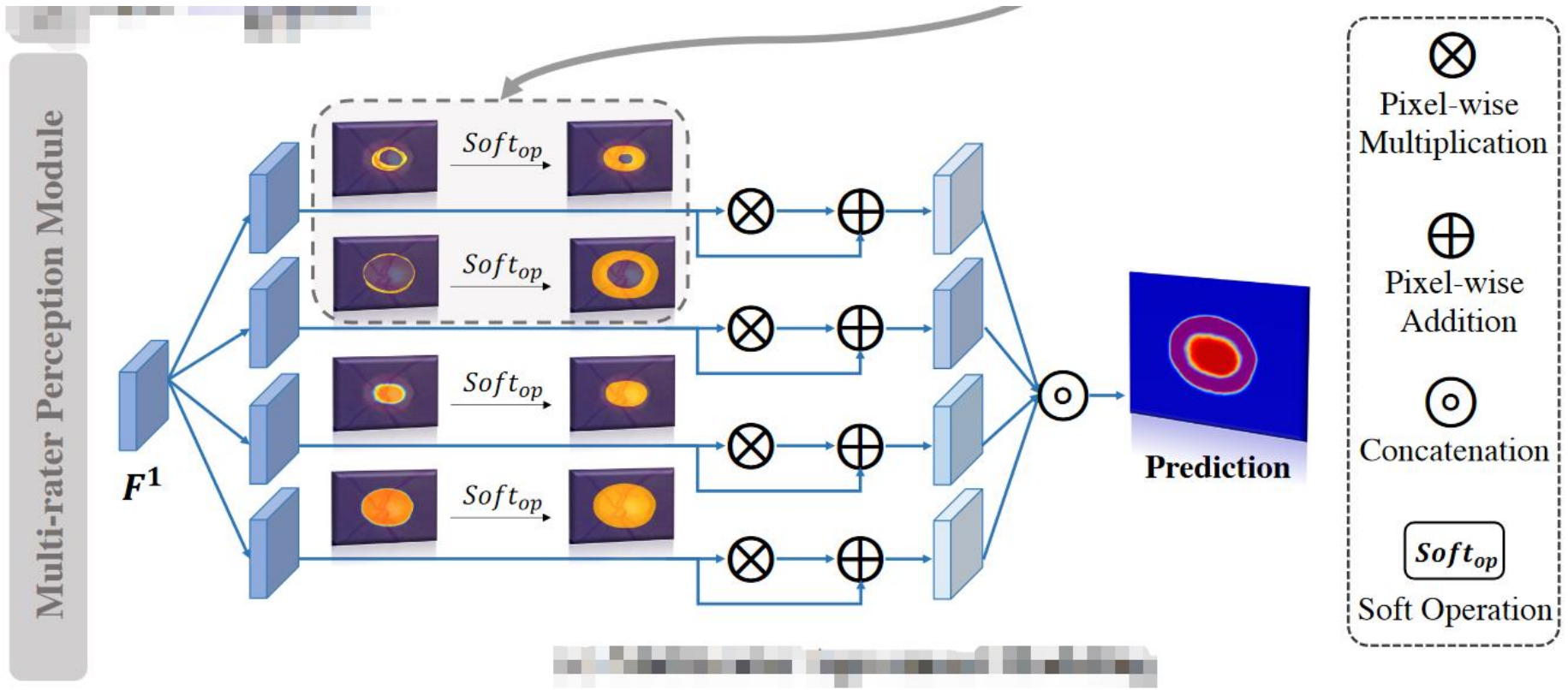


- Calculation of Uncertainty Map:

$$\mathrm{U_{map}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\bar{S}_i - \frac{1}{N}\sum_{i=1}^{N}\bar{S}_i\right)^2}$$

Architecture

MPM



Architecture

MPM

- Soft Operation : $Soft(U_{\text{map}}) = \Omega_{\max}(\mathcal{F}_{\text{Gauss}}(U_{\text{map}}, k), U_{\text{map}}),$
- Input: F^1 & A_j ---> Joint task for cup and disc segmentation

$$\mathcal{A}_j = \{U^{\text{cup}}_{\text{map}}, U^{\text{disc}}_{\text{map}}, P^1_{\text{cup}}, P^1_{\text{disc}}\}^4_{j=1}$$

- Procedure of Soft Attention----> Output M

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

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

- Loss: BCE Loss between (M, GT^{Soft})

Architecture

Final Loss:

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

Alpha = 0.7

Source Codes are publicly available.

According to my observation, loss and optimizer settings are different in the codes, which confuses me.

Dataset

1. RIGA benchmark dataset[6]:
   - Retinal fundus images with optic disc and optic cup annotations for glaucoma analysis.
   - http://www.fao.org/economic/riga/riga-database/riga-request/en/
   - Contains totally 750 color fundus images from three sources, including:
     1. 460 images from MESSIDOR
     2. 195 images from BinRushed
     3. 95 images from Magrabia
   - 6 glaucoma experts fromdifferent organizations labeled the optic cup and disc con-tour masks manually for the RIGA benchmark
   - 195 samples from BinRushed and 460 samples from MESSIDOR were selected as the training set;
   - 95 samples from Magrabia are selected as the test set to evaluate the model, which is not homologous to the training dataset

Dataset

2. QUBIQ [7] :
   - Quantification of Un-certainties 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:
     1. brain growth (one task, MRI, seven raters, 34cases for training and 5 cases for testing),
     2. brain tumor (onetask, MRI, three raters, 28 cases for training and 4 cases fortesting),
     3. prostate (two subtasks, MRI, six raters, 33 casesfor training and 15 cases for testing),
     4. kidney (one task, CT,three raters, 20 cases for training and 4 cases for testing)

Results

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**)

MV ---- Majority Vote

LS ----- Label Sample

MH ---- Multi Heads

Results

- RIGA

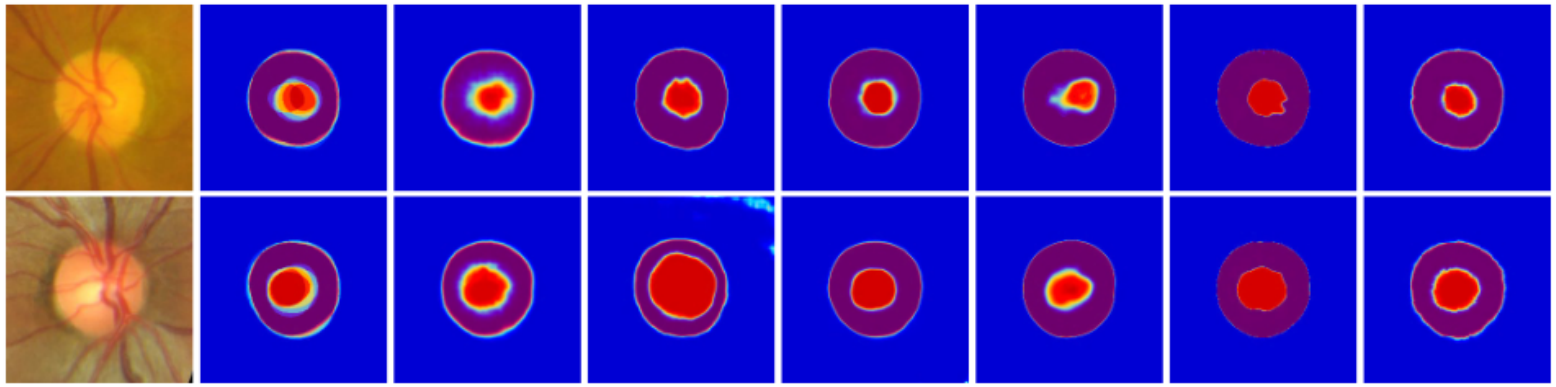


Figure 4. Visual comparisons of our MRNet with the state-of-the-arts for joint optic cup and disc segmentation tasks.

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

	$\mathcal{D}^s_{disc}$ (%)	$\mathcal{D}^s_{cup}$ (%)	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**

Ablation Study

- RIGA dataset

Table 4. Ablation analysis on the RIGA test set.

Module						Average *Expertness*	
Index	Baseline	EIM	ConvLSTM	MRM	MPM	$\mathcal{D}^{s}_{disc}$ (%)	$\mathcal{D}^{s}_{cup}$ (%)
(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)	$loss_{rec}$	$loss_{con}$	MPM	$Soft_{op}$	$\mathcal{D}^{s}_{disc}$ (%)	$\mathcal{D}^{s}_{cup}$ (%)
(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

Results

- QUBIQ dataset

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

(%)	$\mathcal{D}^{s}_{\text{kidney}}$	$\mathcal{D}^{s}_{\text{brain}}$	$\mathcal{D}^{s}_{\text{tumor}}$	$\mathcal{D}^{s}_{\text{pros1}}$	$\mathcal{D}^{s}_{\text{pros2}}$
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**

Results

- QUBIQ dataset

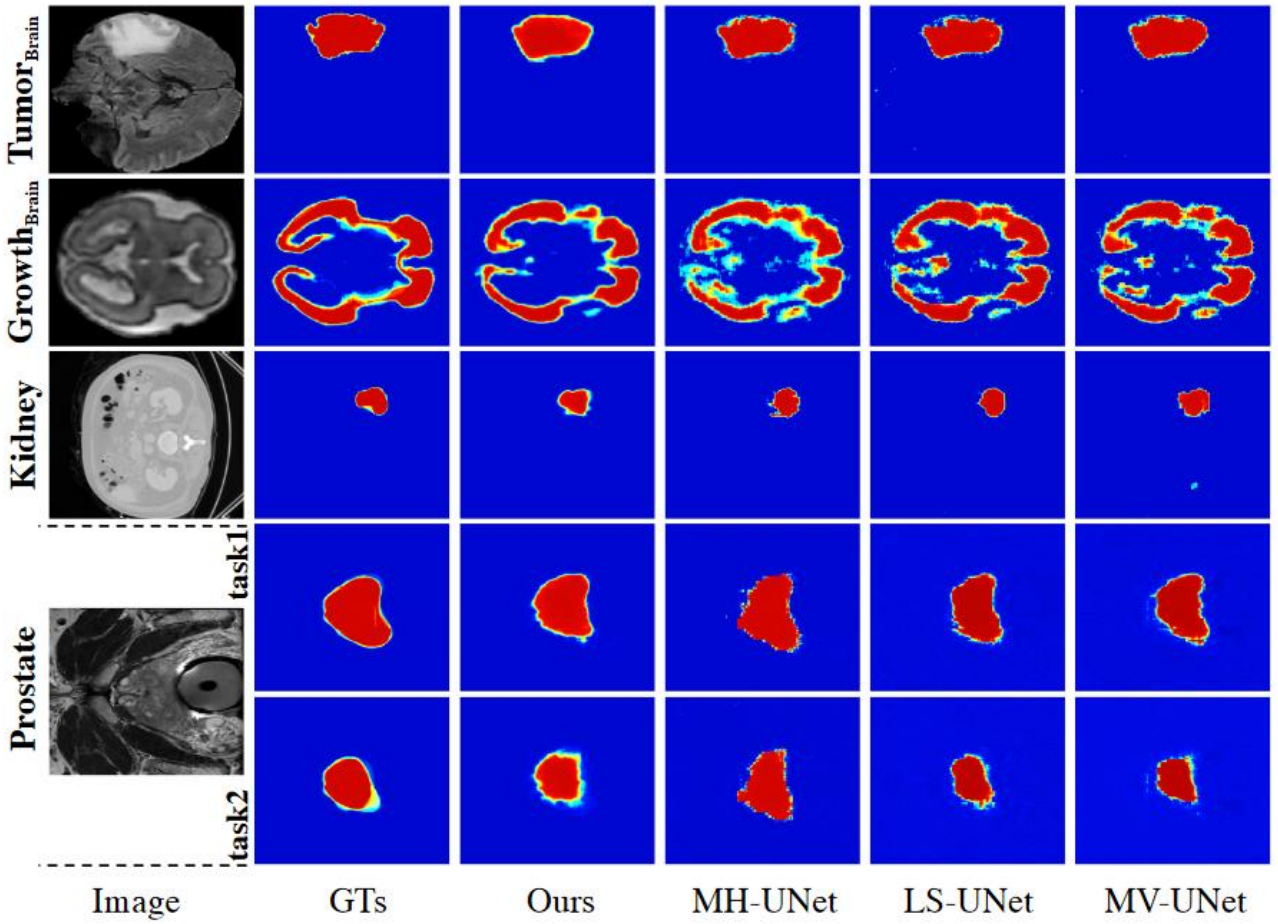


Figure 5. Segmentation results of different strategies for four different medical segmentation tasks on the QUBIQ dataset.

References

1. 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. Christian F Baumgartner, Kerem C Tezcan, Krishna Chai-tanya, Andreas M H ̈otker, Urs J Muehlematter, Khoschy Schawkat, Anton S Becker, Olivio Donati, and Ender Konukoglu. PHiSeg: Capturing uncertainty in medical im-age segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 119–127. Springer, 2019.

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

4. Melody Guan, Varun Gulshan, Andrew Dai, and Geoff Hin-ton. Who said what: Modeling individual labelers improves classification. In *AAAI Conference on Artificial Intelligence*, pages 3109–3118, 2018.

5. 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.

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

7. Bjoern Menze, Leo Joskowicz, Spyridon Bakas, An-dras 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.

QUBIQ Challenge
Miccai 2020 - 2021

QUBIQ

- Quantification of Un-certainties in Biomedical Image Quantification Challenge, is a recently available challenge dataset specifically for the evaluation of inter-rater variability
- https://qubiq.grand-challenge.org/
- QUBIQ contains four different segmentation datasets with CT and MRI modalities, including:
    1. brain growth (one task, MRI, seven raters, 34cases for training and 5 cases for testing),
    2. brain tumor (one task, MRI, three raters, 28 cases for training and 4 cases for testing),
    3. prostate (two subtasks, MRI, six raters, 33 cases for training and 15 cases for testing),
    4. kidney (one task, CT, three raters, 20 cases for training and 4 cases for testing)

2020 Methods

- Wei. Ji & Wenting. Chen et. al
- Uncertainty Quantification for Medical Image Segmentation Using Dynamic Label Factor Allocation Among Multiple Raters (0.77)

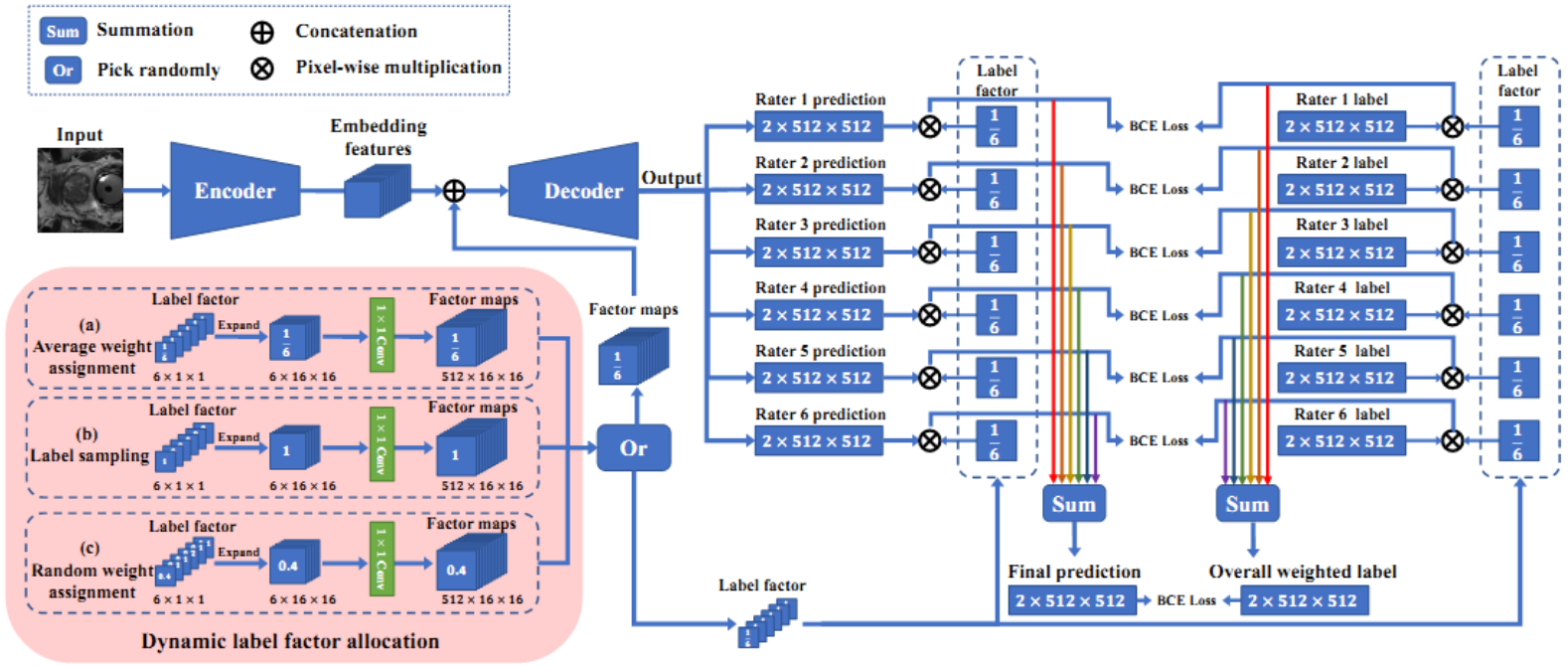


Fig. 1. Architecture of our proposed framework, including a segmentation network based on U-Net, the dynamic label factor allocation mechanism and the training loss for each rater.

2020 Methods

- Jun. Ma
- Estimating Segmentation Uncertainties Like Radiologists

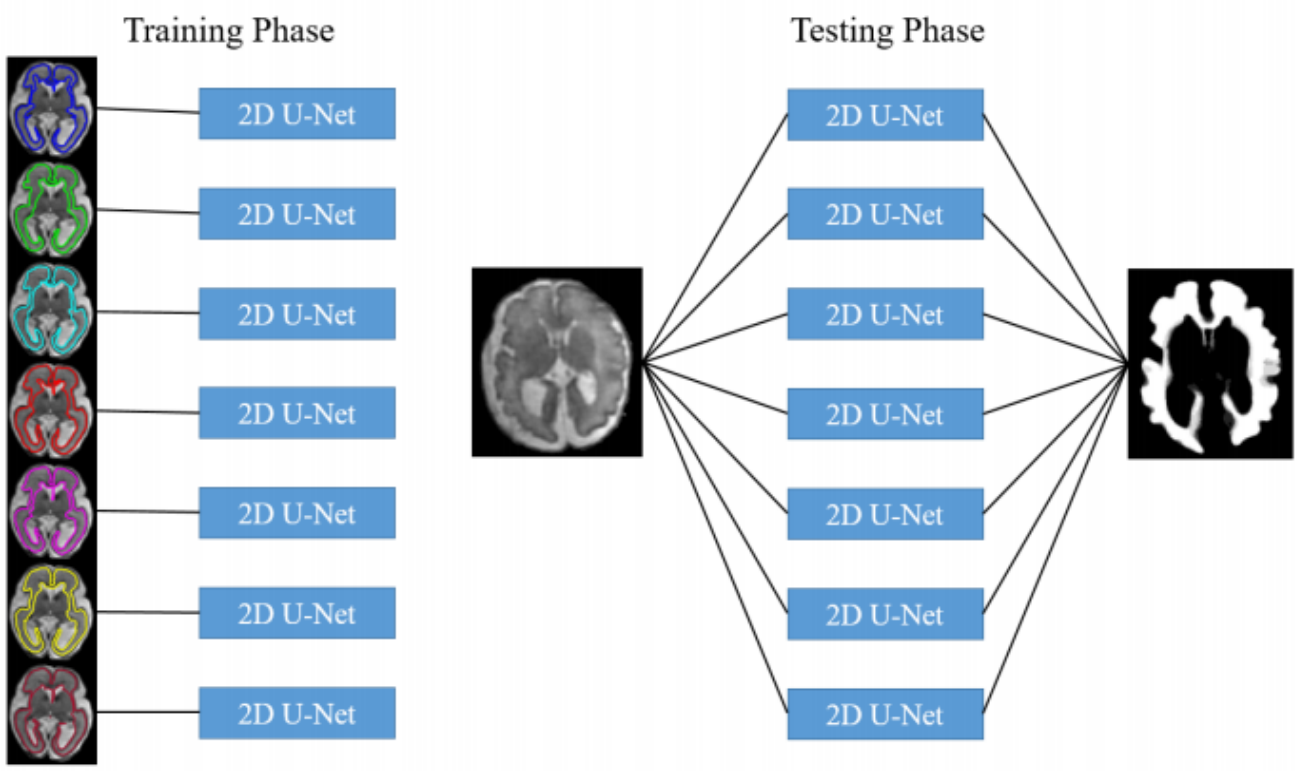


Fig. 2. Training and testing phases of the proposed uncertainty estimation methods.

2020 Methods

- Yanwu Yang and Ting Ma
- Integrated Segmentation with multiple annotations

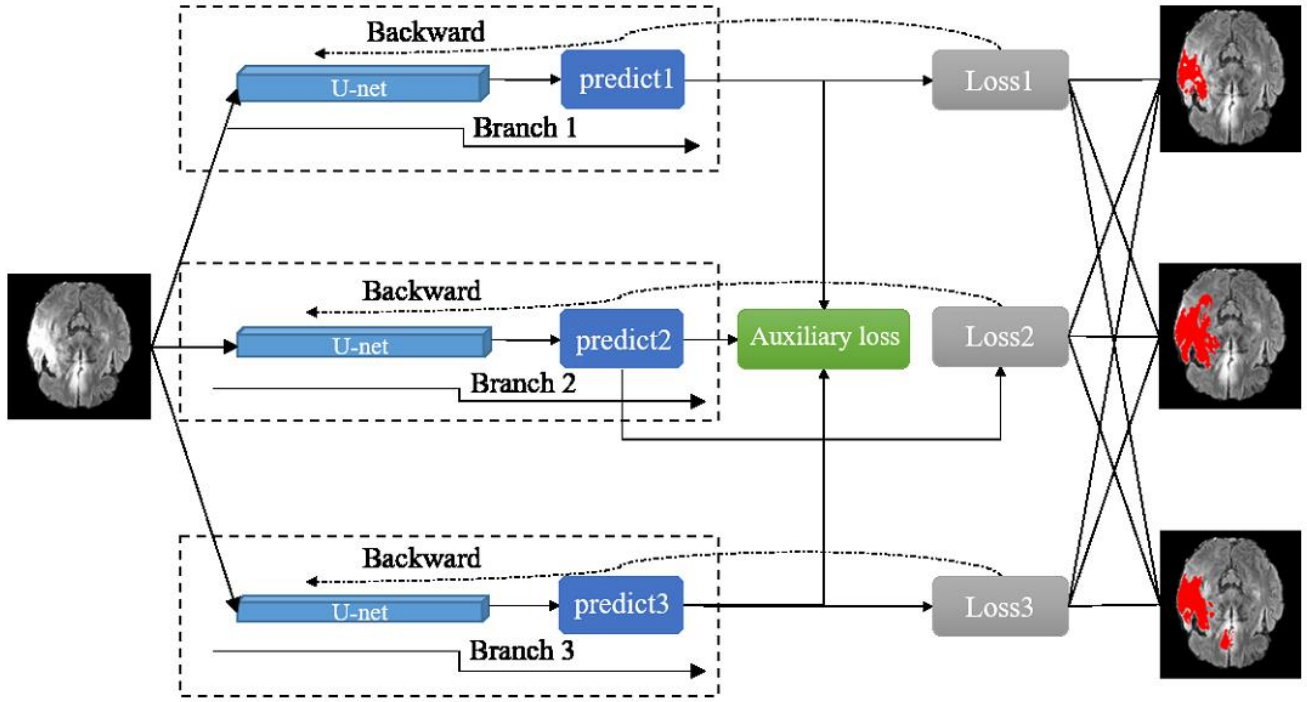


Fig. 2. Integrated architecture designed for multiple annotations, with an auxiliary loss and a loss in each branch.