Science Bulletin 69 (2024) 2906–2919

Contents lists available at ScienceDirect

Science Bulletin

journal homepage: www.elsevier.com/locate/scib

Article

Multi-rater Prism: Learning self-calibrated medical image segmentation from multiple raters

Junde Wu [a,b,c,1], Huihui Fang [a,b,d,1], Jiayuan Zhu [c], Yu Zhang [e], Xiang Li [f], Yuanpei Liu [g], Huiying Liu [h], Yueming Jin [i], Weimin Huang [h], Qi Liu [a], Cen Chen [a], Yanfei Liu [d], Lixin Duan [f,j], Yanwu Xu [a,b,*], Li Xiao [j,*], Weihua Yang [k,*], Yue Liu [d,*]

[a] School of Future Technology, South China University of Technology, Guangzhou 511442, China
[b] Pazhou Lab, Guangzhou 510320, China
[c] The University of Oxford, Oxford OX14AL, UK
[d] Cardiovascular Disease Center, Xiyuan Hospital of China Academy of Chinese Medical Sciences, Beijing 100091, China
[e] State Key Laboratory of Pulsed Power Laser Technology, College of Electronic Engineering, National University of Defense Technology, Hefei 230037, China
[f] Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen 518110, China
[g] The University of Hong Kong, Hong Kong 999077, China
[h] Institute for Infocomm Research, A*STAR, Singapore 138632, Singapore
[i] National University of Singapore, Singapore 119276, Singapore
[j] Sichuan Provincial People's Hospital, University of Electronic Science and Technology of China, Chengdu 611731, China
[k] Shenzhen Eye Hospital, Jinan University, Shenzhen 518040, China

ARTICLE INFO

Article history:
Received 1 September 2023
Received in revised form 3 June 2024
Accepted 7 June 2024
Available online 23 July 2024

Keywords:
Medical image segmentation
Multiple raters
Self-calibration
Half-quadratic algorithm

ABSTRACT

In medical image segmentation, it is often necessary to collect opinions from multiple experts to make the final decision. This clinical routine helps to mitigate individual bias. However, when data is annotated by multiple experts, standard deep learning models are often not applicable. In this paper, we propose a novel neural network framework called Multi-rater Prism (MrPrism) to learn medical image segmentation from multiple labels. Inspired by iterative half-quadratic optimization, MrPrism combines the task of assigning multi-rater confidences and calibrated segmentation in a recurrent manner. During this process, MrPrism learns inter-observer variability while taking into account the image's semantic properties and finally converges to a self-calibrated segmentation result reflecting inter-observer agreement. Specifically, we propose Converging Prism (ConP) and Diverging Prism (DivP) to iteratively process the two tasks. ConP learns calibrated segmentation based on multi-rater confidence maps estimated by DivP, and DivP generates multi-rater confidence maps based on segmentation masks estimated by ConP. Experimental results show that the two tasks can mutually improve each other through this recurrent process. The final converged segmentation result of MrPrism outperforms state-of-the-art (SOTA) methods for a wide range of medical image segmentation tasks. The code is available at https://github.com/WuJunde/MrPrism.

1. Introduction

In clinical practice, it is often necessary to integrate opinions from multiple experts to make the final decision. Therefore, in medical image analysis tasks, it is common to collect the data with multiple labels annotated by different clinical experts. This is particularly relevant in medical image segmentation, where labels are often highly subjective. As an example, Fig. 1a shows the segmentation of the optic cup in fundus images, where there is a large variance between different annotations [1]. This makes it difficult to directly apply deep learning models that work well on natural images to this scenario. This problem has been referred to as the "multi-rater problem" in prior works [2–4].

Learning from multi-rater labels, or labels from crowds has a long history in the literature, dating back to the work of Dawid and Skene [5]. Following their method, a number of papers have extended the generative model under certain assumptions to handle various scenarios [6–8]. In the past two decades, Exceptation Maximization based methods [9–14] have become popular.

* Corresponding authors.
E-mail addresses: ywxu@ieee.org (Y. Xu), xiao1985621@163.com (L. Xiao), benben0606@139.com (W. Yang), liuyueheart@hotmail.com (Y. Liu).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.scib.2024.06.037
2095-9273/

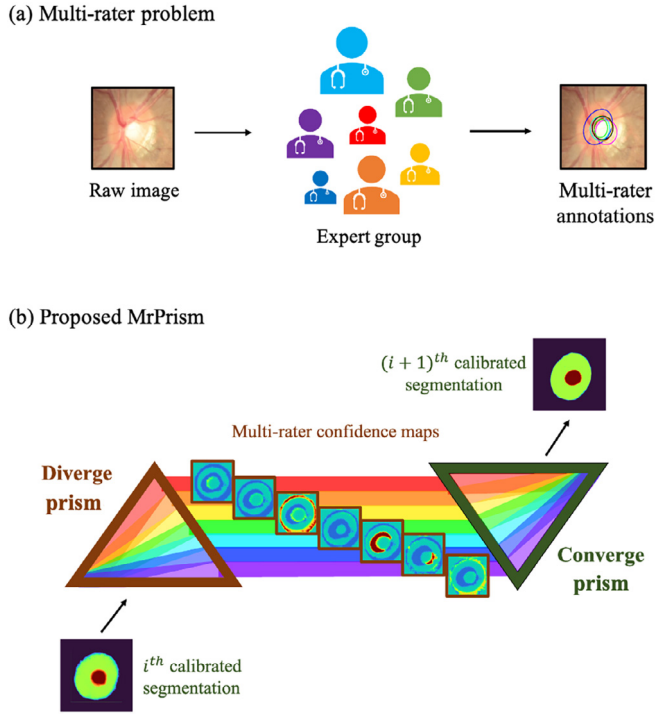


Fig. 1. Multi-rater problem and the proposed MrPrism framework. (a) A multi-rater example of optic-cup annotation, where we can observe large inner-variance among the annotations. The raw fundus image shown here comes from the public dataset REFUGE [1]. (b) In the proposed MrPrism solution, Diverge Prism estimates the multi-rater confidences from the calibrated segmentation mask provided by Converge Prism. Converge Prism again predicts the calibrated segmentation mask from the multi-rater confidences provided by Diverge Prism. Through such an iterative optimization, the segmentation results can be gradually calibrated and refined.

Warfield et al. [9] proposed Simultaneous Truth and Performance Level Estimation (STAPLE) based on EM framework. Then Asman et al. [13] proposed COLLATE method, which is based on STAPLE to add pixel consistency. And, Asman et al. [14] also proposed an extension to the STAPLE approach to seamlessly account for spatially varying performance by extending the performance level parameters to account for a smooth, voxel-wise performance level field unique to each rater. The basic idea is to estimate the posterior probability of each rater by taking the classifier predictions as the prior. Rodrigues et al. [10] extended this idea to neural network-based classifiers to facilitate learning. However, these methods learn fused labels without calibration. As a result, the multi-rater confidences cannot be dynamically adjusted during the inference stage, leading to either overconfident or ambiguous results. For example, when the confidences are implicitly learned, the results tend to be overconfident [10,15]. Conversely, when the confidences are explicitly learned, the results tend to be ambiguous [11,12].

Another branch of study focused on capturing the inter-rater variety by learning from the multi-rater labels [3,16–19]. A range of techniques have been proposed, including Monte Carlo dropout [20,21], ensembles [22,23], multi-head networks [16,18], and variational Bayesian inference [24–26]. One approach is to adopt a label sampling strategy to sample labels randomly from the multi-rater labeling pool during training [19,27]. This results in better calibration than training on the traditional majority vote. Another approach is to model each rater individually by multiple decoders in a neural network [16,17], showing that the multi-head supervision outperforms that of a unique combined ground-truth. Ji et al. [3] proposed a calibrated model to estimate the corresponding segmentation masks based on any given multi-rater confidence, achieving remarkable performance on various fused labels. However, it still requires users to provide the correct multi-rater confidence in advance, which limits its application.

We note that on the one hand, calibrated segmentation requires knowledge of the raters' confidence in order to predict the segmentation mask accurately. On the other hand, multi-rater confidence assignment lacks a calibrated segmentation model to improve its dynamic representation capability. To address the shortcomings of these two research branches, we propose a novel framework called Multi-rater Prisms (MrPrism) that jointly learns calibrated segmentation and multi-rater confidence assignment in a recurrent manner (Fig. 1b). Through this approach, the two tasks can be optimized together to achieve mutual improvement. Ultimately, our proposed recurrent learning technique can produce a self-calibrated segmentation mask consistent with inter-observer agreement.

The proposed recurrence process follows the framework of iterative half-quadratic optimization, which simultaneously optimizes calibrated segmentation and multi-rater confidence assignment, while also considering the raw image's structural prior as a constraint. Specifically, our MrPrism consists of two components: Converging Prism (ConP) and Diverging Prism (DivP). ConP learns calibrated segmentation based on the multi-rater confidences provided by DivP. It takes raw image segmentation features as input and uses the multi-rater confidences as conditions to predict a unique segmentation mask, which allows the neural network to implicitly learn the raw image structural prior. On the other hand, DivP learns to estimate multi-rater labels from the segmentation masks provided by ConP. The segmentation probability maps produced by DivP can naturally represent the raters' confidences (as shown in Section 2.2). Through the iterative optimization of DivP and ConP, both calibrated segmentation and multi-rater confidences can gradually converge to the optimal solutions under the constraint of the raw image structural prior.

ConP and DivP are mainly implemented using the self-attention mechanism [28] to capture the tissue/organ structure globally. ConP is implemented using a combination of dot-product attention and deconvolution layers. Thanks to the global and dynamic nature of the attention mechanism, ConP can naturally integrate the dynamic representations of multi-rater confidences with segmentation features. DivP is implemented using multi-head dot-product attention, with each head estimating the corresponding rater annotation. The estimated probability maps are then used to represent the multi-rater confidences. To avoid trivial solutions, we also shuffle the multi-head loss function of DivP to disentangle the multi-rater segmentation maps and confidences. The experimental results demonstrate that the proposed method outperforms previous state-of-the-art methods on various medical image segmentation tasks. The preliminary work of this paper was previously published at MICCAI 2022. This paper extends the conference paper, providing detailed theoretical derivations, designing a new shuffle supervision mechanism, supplementing experiments on a richer set of image modalities, and including comparative experiments with the latest methods. The extended content is more than 60%.

In brief, this paper's four main contributions are as follows:

(1) We propose MrPrism, a novel medical image segmentation framework that learns from multi-rater annotated labels. Guided by the iterative half-quadratic optimization, MrPrism can calibrate the segmentation by itself through recurrent learning.

(2) In MrPrism, we design ConP and DivP to learn the calibrated segmentation and multi-rater confidences assignment in a recurrent manner, achieving mutual improvement on both tasks.

(3) We carefully design the structure and supervision of MrPrism for iterative optimization. The proposed techniques, including attentive feature integration and shuffled supervision, are found to be effective for recurrent learning.

(4) We conduct comprehensive experiments on a wide range of medical segmentation tasks. The experimental results show that MrPrism consistently outperforms previous state-of-the-art strategies. The improvement is particularly significant when inter-rater variability is large, demonstrating the superior ability of MrPrism to handle multi-rater disagreements.

2. Materials and methods

2.1. Datasets

This study is conducted on five public datasets: REFUGE [1], RIGA [29], QU-BraTS [30], QUBIQ [31], and LIDC-IDRI [32], each representing different types of medical segmentation tasks. For example, the REFUGE dataset is designed for the segmentation of the optic cup and disc in fundus photographs. Since all data used in this study are from public datasets, an exemption from ethical review was obtained from Xiyuan Hospital of China Academy of Chinese Medical Sciences.

2.1.1. REFUGE

The experiments of OD/OC segmentation from fundus images are conducted on two publicly released benchmarks, REFUGE and RIGA. REFUGE [1] is a publicly available dataset for optic cup and disc (OD/OC) segmentation and glaucoma classification. It contains a total of 1200 color fundus images, including three sets each with 400 images for training, validation, and testing. Seven glaucoma experts from different organizations first labeled the optic cup and disc contour masks manually, and a senior expert along with the seven graders, arbitrated the final ground-truths for the validation and testing. Among these, 120 samples correspond to glaucomatous subjects, and the others correspond to non-glaucomatous subjects. The glaucomatous subjects are distributed equally to the training, validation, and test set.

2.1.2. RIGA

RIGA benchmark [29] is a publicly available dataset for OD/OC segmentation, which contains 750 color fundus images in total. Six glaucoma experts from different organizations labeled the optic cup and disc contour masks manually. We select 655 samples in it for the training and 95 heterologous samples for the testing. A senior glaucoma expert with more than ten years of experience was invited to arbitrate the annotations of the test set.

2.1.3. QU-BraTS

The brain tumor segmentation from Magnetic Resonance Imaging (MRI) images is conducted on QU-BraTS 2020 [30] dataset. The dataset contains 341 cases for training, 153 cases for validation, and 166 cases for testing. The cases are annotated by two to four raters and arbitrated by two senior clinical experts.

2.1.4. QUBIQ

QUBIQ benchmark [31] is a recently available challenge dataset specifically for the evaluation of inter-rater variability. QUBIQ contains four different segmentation datasets with Computed Tomography (CT) and MRI modalities. Experts in relevant fields are invited to arbitrate the annotations of the validation and test samples respectively. The tasks include brain growth (MRI, seven raters, 34 cases for training and 5 cases for testing), brain tumor (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), and kidney (CT, three raters, 20 cases for training and 4 cases for testing).

2.1.5. LIDC-IDRI

LIDC-IDRI [32], is an open medical imaging database primarily used for computer-aided diagnosis and research of the lungs. The dataset obtained through the official pip package contains 1010 patients, each with 8 slices. Each slice has a lung nodule segmentation annotated by four experienced thoracic radiologists and is not labeled if a radiologist thought there is no lung nodule here. In the experiments, we split the dataset into a training set composed of 722 patients, a validation set composed of 144 patients, and a test set composed of the remaining 144 patients according to [25].

2.2. Theoretical analysis

In this section, we formally define the problem and give the theoretical premises of MrPrism. Suppose that there are M raters, and K classes, e.g., optic disc, optic cup, background in optic cup/disc segmentation task. Let $z^m \in R^{H\times W\times K}$ denote the label that rater m assigns to an item, where H and W are the item's height and width, respectively. Let $z^{[M]}$ denote the set of all labels assigned by the M raters, $z^{[M]} = \{z^1, z^2 \cdots z^M\}$. We assume that the data points X and the multi-rater labels $z^1, z^2 \cdots z^M$ are drawn independently and identically distributed from the random variables X and $Z^1, Z^2 \cdots Z^M$.

We denote by y the fusion of the M raters' labels $z^{[M]}$ by multi-rater confidence maps $w^{[M]}$, expressed as follows:

$$y = \mathrm{softmax}(\sum_{m=1}^{M} w^m \cdot z^m + p_u), \tag{1}$$

where $\cdot$ represents element-wise multiplication of two matrices, p_u is the uniform prior (a matrix with constant values), w^m represents the confidence maps of rater m and $w^{[M]} = \{w^1, w^2 \cdots w^m\}$. We use the softmax function to normalize the matrix dimension that represents the classes to ensure that the sum of the possibilities is one.

Then, we can formally model the issue as:

$$argmin_{W,Y} ||W \times Z - Y||_2^2 + \eta P_x, \tag{2}$$

where W represents the optimization variables w.r.t the multi-rater confidence maps $w^{[M]}$ and prior p_u, Y represents the calibrated segmentation ground-truth which can be fully represented by the optimal confidence maps and prior according to Equation (1), Z denotes the observed multi-rater label matrix $z^{[M]}$, The $\times$ symbol implies general weighted sum operation, which is defined as $\mathscr{W} \times \mathscr{Z} = \mathrm{softmax}\left(\sum_{m=1}^{M} w^m \cdot z^m + p_u\right)$. P_x is the constraint image prior related to the raw image x, η is a weighting constant. Our goal is to estimate the W together with the calibrated mask Y, which minimizes Eqn. (2) when given multi-rater labels Z and raw image x.

Directly solving Eqn. (2) is difficult since both terms contain unknown optimization variables. However, the problem can be simplified using the iterative half-quadratic optimization method

[33]. With half-quadratic minimization, Eqn. (2) is equivalent to the iterative optimization of the following equations:

$$\begin{cases} W_i' = \quad argmin_{W_i'} \frac{\beta}{2}\|W_{i-1} - W_i'\|_2^2 + \eta P_x(W_i') & ① \\ W_i = \quad argmin_{W_i} \frac{\beta}{2}\|W_i - W_i'\|_2^2 + \frac{1}{2}\|W_i \times Z - Y_i\|_2^2 & ②, \end{cases} \tag{3}$$

where W_i' is an auxiliary variable introduced to relate two equations, i is the number of iterations, β is a variable parameter and Y_i represents the calibrated segmentation mask in i_{th} iteration. Eqn. (3) can be solved by alternatively solving two sub-problems with increasing β.

Such an observation inspired us to design two sub-models to fit the two sub-problems respectively and run iteratively to solve the final optimization problem. In particular, we design ConP to fit the sub-problem (3)-① and DivP to fit the sub-problem (3)-②. To conceptually understand the design, we can see the first terms of Eqn. (3)-① and (3)-② optimize their current optimizing variables to get close to the other's last predicted result. We fulfill this condition in MrPrism by supervising the two sub-models with each other's last predicted result in the training stage. The second term of Eqn. (3)-① is a constraint of image prior. We make it self-learned implicitly by inputting the raw images into the neural network-based sub-model (ConP). Because of the continuous nature of the neural network (similar inputs in different image block locations tend to get similar outputs) [34–36], the network is prone to output the segmentation mask following the raw structure of the image.

The second term of Eqn. (3)-② encourages the optimal multi-rater confidences W_i given the multi-rater labels Z and correctly fused mask Y_i. However, we know neither the optimal multi-rater confidence nor the correctly fused mask, which makes supervision hard. To address the issue, we find a relationship between them, allowing us to directly supervise the sub-model through the multi-rater labels, which can be represented as:

Proposition 1. The confidence map w^m can be obtained from the natural logarithm of the estimated probability map of the label z^m given the fused mask. That is, $w^m = logP(z^m|y)$.

The proof of Proposition 1 is put in the Supplementary materials. Based on Proposition 1, we can supervise DivP through the multi-rater labels $z^{[M]}$ and take its output $\tilde{z}^{[M]} = P(z^{[M]} \mid y)$ as the multi-rater confidence maps. In this way, the constraint in Eqn. (3)-② can be constructed by the supervision of the fused masks, as the loss function $\mathscr{L}_{sff}$ we designed in Section 2.3.3.

To formally describe this work flow, we first define two kinds of fused masks according to Proposition 1, which will be used widely in the later discussion.

Definition 1.

$$y^{fuse} \doteq \tilde{z}^{[M]} \odot z^{[M]} \doteq softmax(\sum_{m=1}^{M} log\left(\tilde{z}^m\right) \cdot z^m + p_u), \tag{4}$$

$$y^{self} \doteq \tilde{z}^{[M]} \odot z^{[M]} \doteq softmax(\sum_{m=1}^{M} log\left(\tilde{z}^m\right) \cdot T(z^m) + p_u), \tag{5}$$

where $\doteq$ means definition, T denotes the thresholding, $\odot$ is the fusion operation, and $\tilde{z}^{[M]}$ represents multi-rater segmentation masks produced by the DivP. Then we can transfer Eqn. (3) to a format more convenient for the implementation:

Proposition 2. Solving Eqn. (3) is equivalent to solving the following equation:

$$\begin{cases} Y_i' = \quad \arg\min_{Y_i'} \frac{\beta}{2}\|Y_{i-1}^{self} - Y_i'\|_2^2 + \eta P_x(Y_i') & ① \\ Y_i^{self} = \quad \arg\min_{Y_i^{self}} \frac{\beta+1}{2}\|Y_i^{self} - Y_i'\|_2^2 + \frac{1}{2}\|Y_i^{fuse} - Y_i^{self}\|_2^2 & ②. \end{cases} \tag{6}$$

Likewise, Y' is an auxiliary variable to relate two sub-equations. The proof of Proposition 2 is provided in the Supplementary materials.

In Eqn. (6), we reorganize all the variables into the same form (multi-masks fusions), which largely facilitates our practical implementation. It not only enables more consistent supervision of the two sub-models, but also allows ConP to only produce the fusions with the confidence maps and image prior implicitly learned.

Specifically, we design ConP as a segmentation decoder that predicts the fusion Y'. ConP network is supervised by Y^{self} produced by DivP in the last iteration (the first term in Eqn. (6)-①). The raw image prior is implicitly learned from the inputted raw image (the second term in Eqn. (6)-①). We design DivP as a multi-rater masks predictor that estimates the multi-rater confidences W based on ConP-predicted segmentation masks Y' in the last iteration. Instead of the direct supervision of the confidences, we supervise the self-fusion of the confidences, i.e., Y^{self} with the last fusion estimated by ConP, i.e., Y' (the first term of Eqn. (6)-②) and the fusion with multi-rater labels, i.e., Y^{fuse} (the second term of Eqn. (6)-②). Such an iterative optimization process will converge to the consistent agreement between ConP and DivP constraint by the raw image structural prior according to the half-quadratic algorithm.

2.3. Methodology

The overall flow of MrPrism is shown in Fig. 2a. Raw image a is first sent into a CNN-based encoder to obtain a deep embedding f_0. Then ConP utilizes f_0 and given confidences $\tilde{w}^{[M]}$ to estimate the calibrated segmentation mask $\tilde{y}$. DivP will then separate $\tilde{y}$ to multi-rater segmentation masks $\tilde{z}^{[M]}$, which represents multi-rater confidence maps $\tilde{w}^{[M]} = \log(\tilde{z}^{[M]})$ according to Proposition 1. $\tilde{w}^{[M]}$ will then be sent to ConP for calibration in the next iteration. ConP and DivP will run recurrently until they converge. We will introduce the implementation of ConP and DivP modules from the methodology level in the following paragraphs. The detailed network structures can be found in the Supplementary materials and released code.

2.3.1. ConP

We propose ConP to estimate calibrated segmentation masks based on the estimated multi-rater confidences. The basic structure is shown in Fig. 2b. The input of ConP is the raw image embedding f_0, and the output is the segmentation mask $\tilde{y}$. Multi-rater confidence maps $\tilde{w}^{[M]}$ estimated by DivP are integrated into ConP through the attentive mechanism to calibrate the segmentation. In ConP, attention is inserted into each two of the deconvolution layers. It takes embedded multi-rater confidence as *query* and segmentation features as *key* and *value*. In this way, the segmentation features can be selected and enhanced based on the given multi-rate confidence maps.

Specifically, consider ConP at the n^{th} layer, the segmentation feature is $f_n \in \mathbb{R}^{\frac{H}{r_n} \times \frac{W}{r_n} \times C_n}$, where $(\frac{H}{r_n}, \frac{W}{r_n})$ is the resolution of the feature

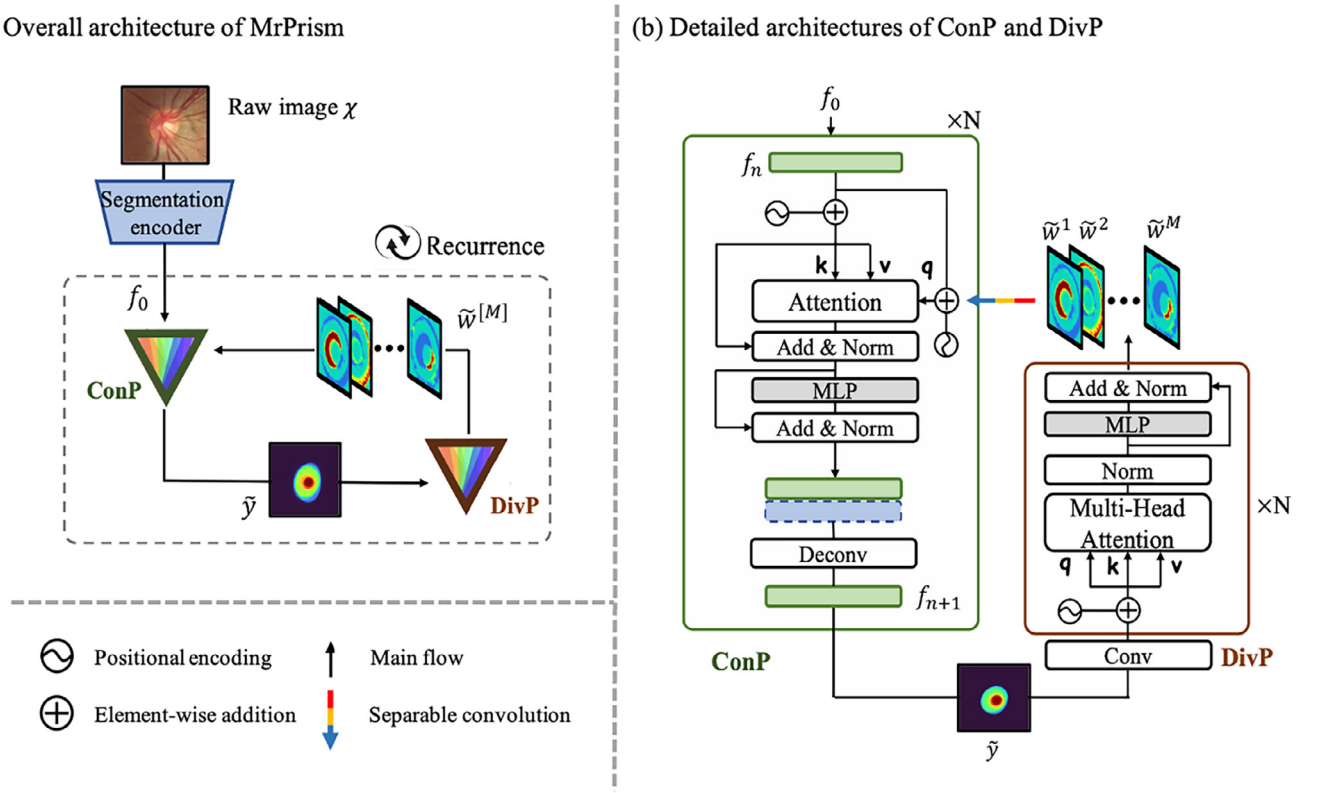


Fig. 2. An illustration of MrPrism framework, which starts from (a) an overview pipeline using optic disc and cup segmentation as an example, and continues with (b) zoomed-in diagrams of individual modules, including ConP and DivP. The raw fundus image shown here comes from the public dataset REFUGE [1].

map and C_n is the number of channels. The embedding of multi-rater confidence maps is $\tilde{w}_n^e \in \mathbb{R}^{\frac{H}{r_n}\times\frac{W}{r_n}\times C_n}$. Then the integration process can be represented as:

$$\bar{f}_n = \text{Attention}(q,k,v) = \text{Attention}(\tilde{w}_n^e + E_w, f_n + E_f, f_n), \tag{7}$$

where $\bar{f}_n$ is the integrated feature, Attention(*query*, *key*, *value*) denotes attention mechanism, E_w, and E_f are positional encodings [37] for confidence embedding and segmentation feature map respectively. Following [38], we reshape the feature maps into a sequence of flattened patches before the attention. Similarly, $\bar{f}_n$ will be reshaped back to $\mathbb{R}^{\frac{H}{r_n}\times\frac{W}{r_n}\times C_n}$ after the attention. $\bar{f}_n$ will then pass through a multi-layer perceptron (MLP) layer to further reinforce the targeted features, The residual connection [39], followed by layer normalization [40], is employed before and after the MLP layer to facilitate the training. The MLP layer consists of two linear mappings that keep the dimension of the input. Then a standard UNet [41] operation is adopted to up-sample the feature. First, it will concatenate with the corresponding encoder feature. Then the deconvolution layer will be applied to obtain $f_{n+1} \in \mathbb{R}^{\frac{H\times 2}{r_n}\times\frac{W\times 2}{r_n}\times\frac{C_n}{2}}$. The blocks are stacked in ConP to achieve the final output $\tilde{y} \in \mathbb{R}^{H\times W\times K}$.

2.3.2. DivP

The basic structure of DivP is shown in Fig. 2b. DivP estimates the segmentation mask of each rater based on the calibrated mask from ConP. The input of DivP is the intermediate calibrated segmentation mask, denoted as $\tilde{y}$. The output of DivP is the estimation of multi-rater segmentation masks, represented as $\tilde{z}^{|M|}$. DivP is implemented using stacked multi-head attention blocks. The final multi-head attention block consists of M heads, with each head estimating one rater's segmentation annotation.

Similar to ConP, MHA (*query, key, value*) denotes multi-head attention mechanism. The multi-head self-attention of the first block in the stack. denoted as $\text{MHA}\left(\tilde{y}+E_y, \tilde{y}+E_y, \tilde{y}\right)$, is first applied on $\tilde{y} \in \mathbb{R}^{H\times W\times K}$, estimated by ConP. E_y are positional encodings. We reshape the feature maps into a sequence of flattened patches before applying attention. Unlike ConP, DivP employs a self-attention strategy, which means *query, key*, and *value* are set as the same when calculating dot-product attention. Residual connection, normalization, and an MLP layer are applied after each head to facilitate training. In DivP, we stack four such blocks. The final estimated multi-rater probability maps are used to calculate the estimated confidences $\tilde{w}^{|M|}$.

In order to integrate the confidences $\tilde{w}^{|M|}$ into ConP attentive blocks, we use separable convolution [42] to embed these maps to the same size as the target segmentation features in ConP. Separable convolution comprises a pair of point-wise convolution and depth-wise convolution for the embedding. Point-wise convolution keeps the scale of the maps but deepens the channels, while depth-wise convolution down-samples the features while retaining the channel number. These layers can not only reshape the maps but can also transfer the maps to deep features for integration, The detailed structure of separable convolution is shown in the Supplementary materials.

2.3.3. Supervision

(1) *Recurrence supervision*: Based on the half-quadratic optimization, the two subproblems are partially constrained by the intermediate results of each other in the iterative process. Following this general guidance, we design the recurrence loss for the supervision of MrPrism. In the recurrence loss, ConP and DivP are supervised by the other's previous predictions. Specifically, con-

sider the supervision of the i^{th} recurrence. The i^{th} output of ConP $\left(\tilde{y}_i\right)$ is supervised by the self-fusion label y_{i-1}^{self} (shown in Eqn. (1)) calculated from the $(i-1)^{th}$ output of DivP. The loss term for the ConP is $L_{ConP}^{i}\left(\tilde{y}_i, y_{i-1}^{\text{self}}\right)$. This restricts ConP from predicting the calibrated segmentation mask following the given multi-rater confidences. The i^{th} output of DivP is supervised by the calibrated segmentation mask $\tilde{y}_i$ provided by the i^{th} ConP. The supervision is imposed on the self-fusion label, i.e., y_i^{self} . Therefore, the loss term for the DivP is $L_{DivP}^{i}\left(y_i^{self}, \tilde{y}_i\right)$. It restricts DivP from predicting the correct multi-rater confidences so that the self-fusion mask can reconstruct the provided calibrated segmentation. An illustration of the loss functions is shown in Fig. 3. Formally, $\mathscr{L}_{\text{rec}}$ for the i^{th} recurrence is represented as:

$$\mathscr{L}_{\text{rec}}^{i} = L_{\text{ConP}}^{i}\left(\tilde{y}_i, y_{i-1}^{\text{self}}\right) + L_{\text{DivP}}^{i}\left(y_i^{\text{self}}, \tilde{y}_i\right). \tag{8}$$

The first term is the supervision of ConP and the second term is that of DivP. Note that ConP is supposed to learn the raw image prior to that cannot be supervised. Therefore, instead of pixel-level supervision, we adopt Structural Similarity Index (SSIM) as the loss function. SSIM restricts the estimated mask from having a similar structure to the self-fusion label but allows for slight differences caused by the raw image prior. Similarly, when calculating DivP loss, we also choose SSIM, that is, $\mathscr{L}_{rec}^{i} = L_{ssim}\left(\tilde{y}_i, y_{i-1}^{self}\right) + L_{ssim}\left(y_i^{self}, \tilde{y}_i\right)$.

(2) *Shuffle the multi-head supervision*: According to Eqn. (3)-①, in addition to the recurrence loss, DivP is also supervised by the multi-rater labels. A basic implementation is to supervise each head of DivP by the corresponding multi-rater label, i.e., multi-head (MH) loss function, which can be represented as:

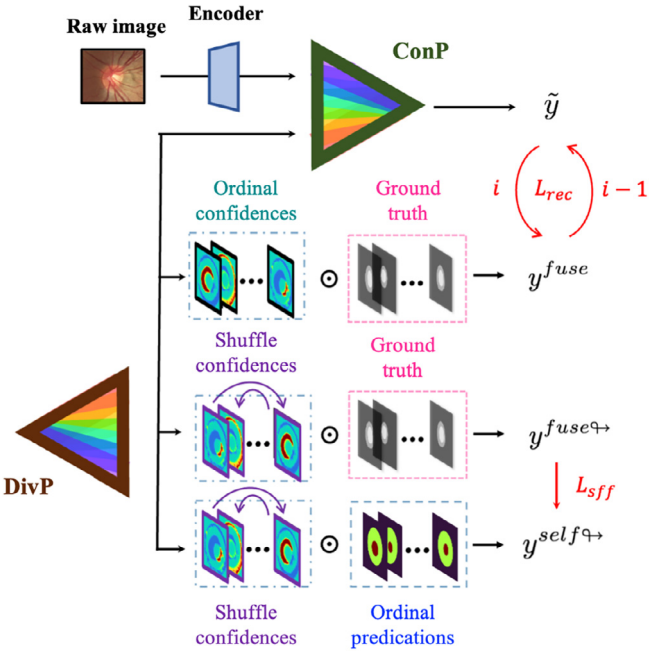


Fig. 3. Supervision of MrPrism. The overall supervision consists of recurrence loss $\mathscr{L}_{\text{rec}}$ and shuffle loss $\mathscr{L}_{\text{sff}}$. The raw fundus image comes from the public dataset REFUGE [1].

$$L_{MH} = \sum_{m=1}^{M} L_{ce}(z^m, z^m), \tag{9}$$

where L_{ce} denotes the cross-entropy loss function. However, since the multi-rater confidence and segmentation are derived from the same estimation, the recurrence is prone to fall into the trivial solution. To avoid this, we disentangle the multi-rater confidences and multi-rater segmentation masks by shuffling. Since under any multi-rater confidences, the fusion of ground-truth labels $z^{[M]}$ and predicated labels $\tilde{z}^{[M]}$ should be the same, the shuffled self-fusions of them can be used for supervision. An illustration of the shuffled supervision is shown in Fig. 3. Specifically, the shuffled supervision of DivP can be represented as:

$$\mathscr{L}_{sff} = L_{ce}\left(y^{selfq}, y^{fuseq}\right), \tag{10}$$

where $\hookrightarrow$ denotes the shuffling on the rater dimension, and

$$\begin{aligned} y^{fuse\hookrightarrow} &= z_i^{[M]\hookrightarrow} \odot z^{[M]} \\ &= soft\max\left(\sum_{m=1}^{M} \log\left(z^{m\hookrightarrow}\right) \cdot z^m + p_u\right), \end{aligned} \tag{11}$$

and,

$$\begin{aligned} y^{self\hookrightarrow} &= \tilde{z}_i^{[M]\hookrightarrow} \odot \mathscr{T}_{0.5}\left(\tilde{z}^{[M]}\right) \\ &= soft\max\left(\sum_{m=1}^{M} \log\left(\tilde{z}^{m\hookrightarrow}\right) \cdot \mathscr{T}_{0.5}\left(\tilde{z}^{m}\right) + p_u\right). \end{aligned} \tag{12}$$

In Eqn. (12), 7.5 denotes the threshold of 0.5. In practice, we empirically shuffle three times in each recurrence for the supervision.

(3) *Overall supervision*: Consider ConP and DivP run once each, i.e., from f_0 to $\tilde{z}^{[M]}$, as one recurrence. Each instance will run τ recurrences in a single epoch. We backpropagate the gradients of the model after τ times of recurrence. The total loss function is represented as:

$$L_{total} = \sum_{i=1}^{\tau} L_{rec}^{i} + \zeta L_{sff}^{i}, \tag{13}$$

where ζ is the weight hyper-parameter. The gradients are back-propagated individually in each recurrence, which means the gradients of the i^{th} recurrence will not affect the $(i-1)^{th}$ recurrence. We show an overall flow of MrPrism algorithm in Algorithm 1.

3. Results

3.1. Experimental setup

3.1.1. Implement details

All the experiments are implemented with the PyTorch platform and trained/tested on 4 Tesla P40 GPU with 24 GB of memory. All training and test images are uniformly resized to the dimension of 256 × 256 pixels. ConP is stacked with five stages with the patch sizes 8, 7, 7, 7, 7 in the attentive mechanism. DivP is stacked with four stages with patch sizes 7, 7, 7, 7, and 4, 4, 4, M heads in the multi-head attention mechanism. The weight constant of shuffle loss ζ is set as 0.3. Generally, each instance will run three times of recurrence in the training and inference stages. The proposed model is trained using Adam optimizer [43] for 150 epochs. The

Algorithm 1. MrPrism Algorithm

Given the segmentation encoder with the parameters θ_e, ConP with the parameters θ_c, and DivP with the parameters θ_d.
Let λ denote the learning rate, ζ denote the weight of shuffle loss, τ denote the number of recurrences. $\mathscr{T}_{0.5}$ denotes the threshold of 0.5, L_{ssim} and L_{ce} denote SSIM and cross-entropy loss functions, respectively.

while *Training* **do**

Sample $(x, z^{[M]})$ from training dataset

Initialize $\tilde{w}_0^{[M]}$ following Eqn. (14)

Initialize $y_0^{f\ \text{use}}$ following Eqn. (15)

$f_0 \leftarrow$ Encoder (x)

for i in range 1 to τ **do** COMMENT# Recurrence process

$\tilde{y}_i \leftarrow \text{ConP}\left(f_0, w_{i-1}^{[M]}\right)$

$\tilde{z}_i^{[M]} \leftarrow \text{DivP}\left(\tilde{y}_i\right)$

$\tilde{w}_i^{[M]} \leftarrow \log \tilde{z}_i^{[M]}$

$y_i^{\text{fuse}} \leftarrow \tilde{z}_i^{[M]} \odot z^{[M]}$

Shuffle $\tilde{z}_i^{[M]}$ to $\tilde{z}_i^{[M]\hookrightarrow}$

$y_i^{\text{fuse}\hookrightarrow} \leftarrow \tilde{z}_i^{[M]\hookrightarrow} \odot z^{[M]}$

$y_i^{\text{self}\hookrightarrow} \leftarrow \tilde{z}_i^{[M]\hookrightarrow} \odot \mathscr{T}_{0.5}\left(\tilde{z}^{[M]}\right)$

$\mathscr{L}_{\text{rec}}^i \leftarrow L_{\text{ssim}}\left(\tilde{y}_i, y_{i-1}^{\text{fuse}}\right) + L_{ssim}\left(y_i^{\text{fuse}}, \tilde{y}_i\right)$

$\mathscr{L}_{sff}^i \leftarrow L_{ce}\left(y_i^{\text{self}\hookrightarrow}, y_i^{\text{fuse}\hookrightarrow}\right)$

end for

$$\theta_e \leftarrow \theta_e + \lambda \sum_{i=1}^{\tau} \frac{\partial L_{rec}^i}{\partial \theta_e}$$

$$\theta_c \leftarrow \theta_c + \lambda \sum_{i=1}^{\tau} \frac{\partial L_{rec}^i}{\partial \theta_c}$$

$$\theta_d \leftarrow \theta_d + \lambda \sum_{i=1}^{\tau} \left(\frac{\partial L_{rec}^i}{\partial \theta_d} + \zeta \frac{\partial L_{sff}^i}{\partial \theta_d}\right)$$

end while

learning rate is always set to 1×10^{-4}. We initialize the recurrence by assigning the equal confidence to each rater. Formally, we set initial w_0^m as:

$$w_0^m = \log[\psi^{\text{m}} + \varepsilon^{\text{m}}]_0^1, \tag{14}$$

where ψ^{m} determines the average of atlas w_0^m, and ε^{m} adds different perturbations to each element. $[\cdot]_0^1$ denotes the value is clipped to range of 0 to 1. For the given w^{m} the ground truth is generated by

$$y_0^{self} = soft\max\left(\sum_{m=1}^{M} w_0^m \cdot z_n^m + p_u\right). \tag{15}$$

We sample $\psi^{[\text{M}]}$ from uniform distribution, $\psi^{[M]} \sim U(a, b)$. ε^{m} is sampled from a normal distribution, $\varepsilon^m \sim \mathcal{N}(\mu, \sigma^2)$, where a, b, μ, σ are set as 0.1, 0.9, 0, 0.2, respectively. The confidence maps will be set as the same for each batch of the data.

3.1.2. Evaluation metric

The segmentation performance is evaluated by soft dice coefficient (D) across 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 [44] is computed. D scores are obtained as the averages of multiple thresholds.

3.2. Experimental results

3.2.1. Overall performance

To verify the self-calibrated segmentation performance of the proposed model, we compare MrPrism with SOTA multi-rater learning methods. The selected methods include WDNet [16], CL [10], CM [15], AggNet [11], MaxMig [12], MRNet [3] and Diag [4]. Specifically, CL, CM, AggNet, MaxMig, and Diag jointly learn the prediction and multi-rater confidence assignment, in which CL, CM, and MaxMig implicitly learn the confidences, while AggNet and Diag explicitly learn the confidences. WDNet and MRNet are calibrated segmentation methods that require rater confidence. Since multi-rater confidences are not available in our scenario, all raters are considered equal when applying these methods. Diag is a method that uses disease diagnosis/classification performance as a standard to evaluate the confidences of multi-rater segmentation labels.

In OD/OC segmentation, we use glaucoma diagnosis as the standard. In the other segmentation tasks, no diagnosis task is associated, so the results of Diag are not reported. The experiments are conducted on a wide range of medical segmentation tasks, including OD/OC segmentation, infant brain growth segmentation, brain tumor segmentation, prostate segmentation, kidney segmentation, and lung nodule segmentation. In MrPrism, the predictions of ConP in four recurrences are used for comparison, denoted as Rec0, Rec1, Rec2, and Rec3. The detailed quantitative results are shown in Table 1.

As listed in Table 1, the proposed MrPrism framework demonstrates superior performance in medical image segmentation tasks with ambiguous annotations compared to SOTA multi-rater learning methods. In addition to OD segmentation task, the results of the proposed method are significantly improved compared with other segmentation tasks. The multi-rater annotation difference of OD segmentation task is small, and the task itself is relatively mature, that is, the segmentation results of each method can easily reach the Dice coefficient of about 95%. The experimental results show that MrPrism can accurately represent the inter-rater agreement, particularly when the inter-rater variety is large, such as in optic cup and brain tumor segmentation tasks. As observed in Table 1, the performance of MrPrism keeps improving through the recurrences on most of the metrics, The most significant improvements are seen in the earlier recurrences, and the results converge after the third recurrence.

We show a visualized comparison of different methods in Fig. 4. We can see CL and CM, which implicitly learn the multi-rater confidences, are inclined to be overconfident in the inaccurate results, while AggNet and MaxMig, which explicitly learn the multi-rater confidences, are prone to obtain ambiguous results. WDNet and MRNet require the correct multi-rater confidences, thus their results tend to be ambiguous under the default majority vote assumption. We show a visualized comparison of different recurrences in Fig. 5. We can see the proposed self-calibrated segmentation can estimate the result from uncertain to confident through the recurrences, finally achieving a confident and calibrated result.

3.2.2. Comparing ConP with SOTA calibrated segmentation methods

The calibrated segmentation methods are commonly proposed to capture the uncertainty from multiple labels by assuming that there is no priority on any of them. Unlike their strategies, we jointly optimize the calibrated segmentation model (ConP) with the priority evaluation (DivP), and achieve stronger calibration ability on ConP. We compare ConP with calibrated segmentation methods, including sPU-Net [24], HPU-Net [26], PHiSeg [25], ICODD [45], DSC++ [46], and CIMD [47] and non-calibrated segmentation methods, including AGNet [48], nnUNet [49], and TransUNet [50] on a wide range of medical segmentation tasks. In the first iteration, we replaced the confidence calculated by the DivP module with the various multi-rater confidence calculated by the

Table 1
Quantitative comparison between MrPrism and SOTA multi-rater learning strategies.

	OD/OC (REFUGE)		OD/OC (RIGA)		Brain growth	Brain tumor	Prostate 1	Prostate 2	Kidney	Lung-nodule
	$\mathscr{D}_{\text{disc}}(P)$	$\mathscr{D}_{\text{cup}}(P)$	$\mathscr{D}_{\text{disc}}(P)$	$\mathscr{D}_{\text{cup}}(P)$	$\mathscr{D}_{\text{brain}}(P)$	$\mathscr{D}_{\text{tumor}}(P)$	$\mathscr{D}_{\text{pros 1}}(P)$	$\mathscr{D}_{\text{pros 2}}(P)$	$\mathscr{D}_{\text{kidney}}(P)$	$\mathscr{D}_{\text{nodule}}(P)$
WDNet [4]	**95.72 (0.15)**	84.16 (1×10^{-5})	96.43 (0.45)	81.55 (5×10^{-4})	83.13 (3×10^{-4})	84.22 (1×10^{-4})	84.62 (7×10^{-4})	73.65 (2×10^{-5})	72.50 (4×10^{-6})	83.12 (5×10^{-4})
CL [19]	95.31 (0.12)	83.41 (2×10^{-5})	95.49 (0.49)	81.11 (2×10^{-4})	81.56 (7×10^{-4})	77.34 (2×10^{-4})	85.59 (3×10^{-4})	73.48 (5×10^{-5})	74.36 (7×10^{-6})	85.87 (3×10^{-4})
AggNet [22]	94.96 (0.08)	83.66 (4×10^{-5})	95.38 (0.57)	82.27 (6×10^{-4})	82.72 (8×10^{-4})	83.37 (4×10^{-4})	78.33 (2×10^{-4})	71.77 (9×10^{-5})	71.39 (1×10^{-5})	84.35 (7×10^{-4})
CM [21]	94.50 (0.13)	84.72 (1×10^{-5})	96.56 (0.60)	83.11 (5×10^{-4})	80.56 (0.001)	79.21 (0.006)	85.91 (2×10^{-5})	74.55 (1×10^{-6})	74.85 (5×10^{-6})	86.05 (1×10^{-4})
MaxMig [20]	**95.72 (0.34)**	84.45 (2×10^{-4})	96.64 (0.62)	84.81 (0.005)	84.62 (5×10^{-4})	86.41 (0.001)	86.27 (2×10^{-4})	74.06 (2×10^{-5})	74.68 (7×10^{-6})	86.60 (3×10^{-4})
MRNet [2]	94.75 (0.43)	85.63 (4×10^{-4})	96.27 (0.55)	85.82 (0.002)	83.67 (0.008)	87.63 (0.012)	87.04 (5×10^{-4})	75.43 (2×10^{-5})	75.19 (1×10^{-5})	87.45 (8×10^{-4})
Diag [3]	95.17 (0.52)	86.23 (0.001)	**96.74 (0.53)**	86.17 (0.007)	–	–	–	–	–	–
MrPrism-Rec0	94.62	84.33	96.18	85.28	83.53	87.81	86.17	75.20	75.24	87.85
MrPrism-Rec1	95.36	87.59	96.37	87.42	84.25	88.23	87.35	76.88	76.22	88.45
MrPrism-Rec2	95.67	**88.74**	96.66	88.15	85.50	**89.58**	**88.45**	77.29	**76.51**	**89.33**
MrPrism-Rec3	95.68	88.53	96.62	**88.30**	**85.52**	89.44	88.37	**77.34**	76.48	88.21

Rec i indicates the results of the i^{th} recurrence, $\mathscr{D}_{xx}$ represents the DICE coefficient of the result of the xx segmentation task, The P value, which is calculated using two-sample independent t-test, reflects the significance analysis of the results of MrPrism and those of other methods, with $P < 0.05$ indicating a significant difference in results. Bold indicates the best results, and '-' indicates that the test was not conducted on this dataset.

majority vote. It is worth noting that all comparison methods use the prior information mentioned above in order to ensure a fair comparison.

The quantitative results are shown in Table 2. As shown in the table, calibrated methods generally outperform non-calibrated ones, and our proposed self-calibrated method performs better than other calibrated methods. Proposed ConP consistently achieves superior performance, indicating its stronger calibration ability. Compared with other calibrated methods, the decomposition process in the recurrence strategy allows ConP to better realize each component of the fusion and ultimately achieve improved calibration results.

3.2.3. Comparing DivP with SOTA label fusion methods

Additionally, the calibrated segmentation of ConP facilitates DivP. We quantitatively compare the self-fusion labels of DivP with SOTA label fusion strategies in Table 3. Methods compared include traditional majority vote (MV), STAPLE [9], COLLATE [13], Spatial STAPLE [14], MaxMig [12], and Diag [4]. In this single-module experiment, we substituted the segmentation ground truth, which should have been produced by the ConP module, for the segmentation mask. Each comparison method also uses the prior information mentioned above to ensure a fair comparison. The quantitative results in Table 3 show that the traditional MV has inferior performance.

STAPLE, an advanced fusion method following the majority first strategy, outperforms the MV method. The COLLATE method introduces pixel consistency on top of STAPLE and thus outperforms STAPLE methods on various tasks. The effect of Spatial STAPLE method is further improved if taking the performance of spatial variations on the basis of STAPLE into consideration. However, these methods do not perform well when few raters are dominant in the annotations, such as in the OC-REFUGE task. MaxMig fuses labels based on raw image structural prior and obtains fair results on most tasks. Diag uses diagnosis labels for label fusion and performs much better, but its application scope is limited, since it requires diagnosis labels. Our proposed method not only fuses labels based on raw image structural prior but can also dynamically adjust the fusions based on calibration, ultimately achieving the best consistency with ground truth.

3.2.4. Ablation study

We conduct detailed ablation study on integration strategies and loss functions. The quantitative results of brain tumor segmentation and OD/OC segmentation are shown in Table 4. We compare our attention-convolution hybrid integration strategy (ConvAtt) with Concat, which concatenates the segmentation and confidence feature for the integration, and ConvLSTM, which uses ConvLSTM [51] for the integration [3].

Thanks to the global and dynamic nature of the scale dot-product attentive mechanism, the proposed integration yields significant improvement compared with ConvLSTM. We then compare the effect of traditional multi-head loss $\mathscr{L}_{MH}$ and the proposed shuffle loss $\mathscr{L}_{ssf}$ as additional constraints for DivP. It is observed that $\mathscr{L}_{ssf}$ outperforms $\mathscr{L}_{MH}$ by disentangling the multi-rater confidences and annotations. Interestingly, applying $\mathscr{L}_{ssf}$ individually works better than applying the combination of $\mathscr{L}_{MH}$ and $\mathscr{L}_{ssf}$. The reason for this outperformance may be that $\mathscr{L}_{MH}$ constrains each head to estimate specific rater annotations, while different separations are possible for one segmentation, so this extra constraint of $\mathscr{L}_{MH}$ eventually degrades the performance. The experimental results indicate that the combination of attentive integration and shuffle loss works best for MrPrism, which is adopted in our final implementation.

An experiment discussing the different loss functions for supervising ConP and DivP modules is also implemented here. Table 5 shows the segmentation effect under rec2 of MrPrism on different segmentation tasks of different datasets when SSIM and Mean Squared Error loss are used for the loss function (Eqn. (8)).

4. Discussion and conclusion

The superior performance of MrPrism on the tasks in Table 1 demonstrates the effectiveness of the proposed framework. It is tailored for medical image segmentation with ambiguous annota-

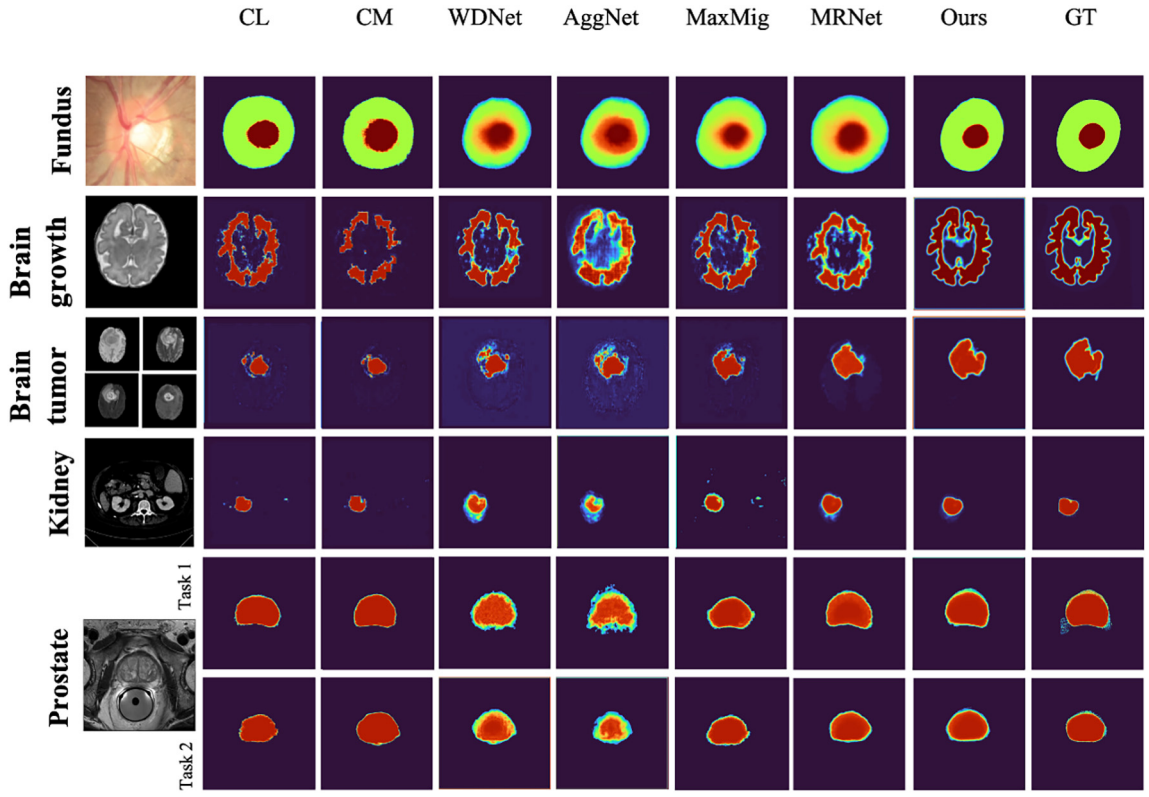


Fig. 4. Visualized comparison with SOTA methods on five different medical segmentation tasks. The raw images used for the task of segmenting the optic cup and disc from fundus images, and of segmenting brain tumors from MRI images are from the public datasets REFUGE [1] and Qu-BraTS [30], respectively. The raw images used for the tasks of segmenting brain growth from MRI images, segmenting kidneys from CT images, and segmenting the prostate from MRI images come from the public dataset QUBIQ [31].

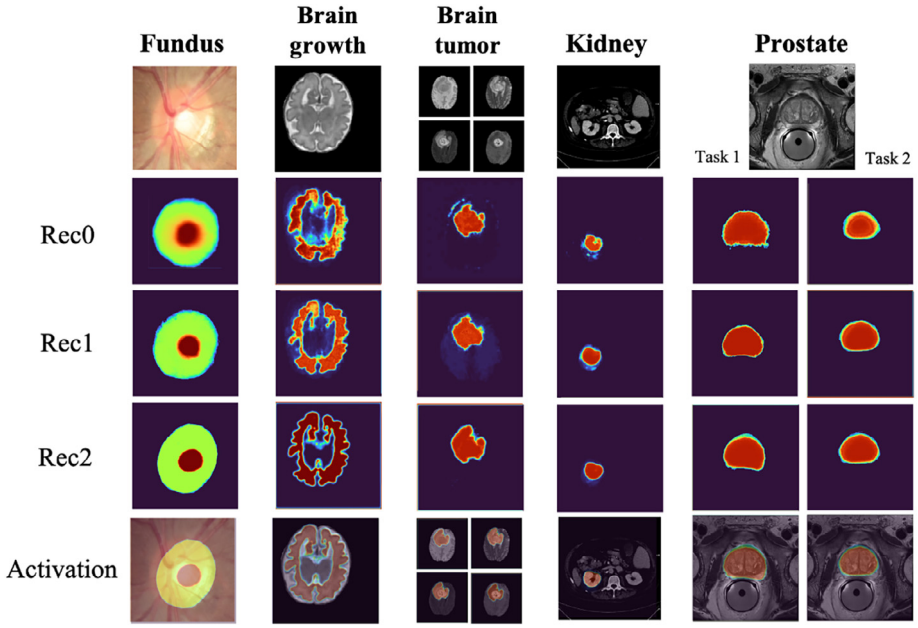


Fig. 5. Visualized comparison of different MrPrism recurrences on five different medical segmentation tasks. The raw images used for the tasks of segmenting the optic cup and disc from fundus images, and of segmenting brain tumors from MRI images are from the public datasets REFUGE [1] and Qu-BraTS [30], respectively. The raw images used for the tasks of segmenting brain growth from MRI images, segmenting kidneys from CT images, and segmenting the prostate from MRI images come from the public dataset QUBIQ [31].

Table 2
Quantitative comparison between ConP and SOTA calibrated/non-calibrated segmentation methods.

		OD/OC (REFUGE)		OD/OC (RIGA)		Brain growth	Brain tumor	Prostate1	Prostate2	Kidney
		$\mathscr{D}_{\text{disc}}$ (P)	$\mathscr{D}_{\text{cup}}$ (P)	$\mathscr{D}_{\text{disc}}$ (P)	$\mathscr{D}_{\text{cup}}$ (P)	$\mathscr{D}_{\text{brain}}$ (P)	$\mathscr{D}_{\text{tumor}}$ (P)	$\mathscr{D}_{\text{pros 1}}(P)$	$\mathscr{D}_{\text{pros 2}}(P)$	$\mathscr{D}_{\text{kidney}}(P)$
No calibrated	AGNet	90.21 (0.38)	71.86 (5×10^{-5})	96.31 (0.23)	78.05 (4×10^{-4})	79.05 (2×10^{-4})	80.38 (6×10^{-4})	84.67 (2×10^{-4})	69.72 (8×10^{-5})	70.69 (1×10^{-5})
	nnUNet	94.72 (0.25)	84.93 (3×10^{-5})	95.91 (0.38)	85.21 (5×10^{-4})	83.51 (4×10^{-5})	86.37 (3×10^{-5})	86.69 (6×10^{-4})	73.75 (4×10^{-5})	73.77 (2×10^{-6})
	TransUNet	94.94 (0.11)	85.61 (7×10^{-5})	97.11 (0.76)	86.17 (3×10^{-4})	83.49 (1×10^{-4})	86.28 (9×10^{-4})	86.80 (3×10^{-4})	74.78 (1×10^{-5})	73.85 (6×10^{-6})
Calibrated	PHiSeg	94.72 (0.20)	85.12 (4×10^{-5})	96.75 (0.69)	86.63 (1×10^{-4})	83.67 (6×10^{-5})	87.44 (4×10^{-5})	86.31 (2×10^{-4})	75.78 (1×10^{-5})	73.25 (8×10^{-6})
	HPU-Net	94.81 (0.17)	85.88 (5×10^{-5})	96.73 (0.53)	86.92 (4×10^{-4})	84.37 (9×10^{-4})	88.26 (8×10^{-5})	87.04 (6×10^{-5})	75.92 (1×10^{-6})	73.89 (8×10^{-5})
	sPU-Net	95.13 (0.26)	86.73 (3×10^{-5})	97.80 (0.44)	87.67 (6×10^{-5})	84.82 (1×10^{-5})	88.95 (6×10^{-5})	87.74 (2×10^{-5})	76.41 (1×10^{-5})	74.58 (3×10^{-6})
	ICODD	95.28 (0.19)	86.91 (6×10^{-5})	97.82 (0.63)	88.01 (4×10^{-5})	84.90 (2×10^{-5})	88.42 (7×10^{-5})	87.21 (4×10^{-5})	76.26 (3×10^{-6})	74.13 (6×10^{-6})
	DSC++	95.65 (0.33)	87.10 (4×10^{-5})	97.47 (0.50)	88.40 (3×10^{-5})	85.39 (7×10^{-4})	88.64 (3×10^{-5})	87.55 (3×10^{-5})	76.69 (2×10^{-5})	75.28 (7×10^{-6})
	CIMD	95.71 (0.49)	87.26 (2×10^{-5})	97.80 (0.86)	88.76 (5×10^{-5})	85.77 (3×10^{-5})	88.80 (6×10^{-5})	87.13 (4×10^{-4})	76.90 (2×10^{-6})	75.33 (5×10^{-6})
Self-calibrated	ConP (ours)	**96.02**	**88.79**	**97.84**	**89.35**	**86.33**	**89.62**	**87.89**	**77.45**	**75.50**

$\mathscr{D}_{xx}$ represents the DICE coefficient of the result of the *xx* segmentation task. The *P* value, which is calculated using two-sample independent *t*-test, reflects the significance analysis of the results of MrPrism and those of other methods, with $P < 0.05$ indicating a significant difference in results. Bold indicates the best results.

Table 3
Quantitative comparison between DivP and SOTA label-fusion methods.

	OD/OC (REFUGE)		OD/OC (RIGA)		Brain growth	Brain tumor	Prostate1	Prostate2	Kidney
	$\mathscr{D}_{\text{disc}}$ (P)	$\mathscr{D}_{\text{cup}}$ (P)	$\mathscr{D}_{\text{disc}}$ (P)	$\mathscr{D}_{\text{cup}}$ (P)	$\mathscr{D}_{\text{brain}}$ (P)	$\mathscr{D}_{\text{tumor}}$ (P)	$\mathscr{D}_{\text{pros 1}}(P)$	$\mathscr{D}_{\text{pros 2}}(P)$	$\mathscr{D}_{\text{kidney}}(P)$
MV	94.52 (0.15)	80.31 (9×10^{-5})	95.60 (0.49)	82.47 (6×10^{-4})	83.06 (7×10^{-4})	79.60 (9×10^{-5})	89.67 (3×10^{-5})	86.10 (3×10^{-5})	84.36 (2×10^{-5})
STAPLE	95.74 (0.10)	80.91 (1×10^{-4})	95.71 (0.57)	83.75 (0.002)	85.41 (0.003)	79.60 (9×10^{-5})	89.38 (1×10^{-4})	86.80 (3×10^{-4})	85.53 (5×10^{-4})
COLLATE	96.01 (0.12)	83.67 (6×10^{-5})	96.72 (0.50)	85.35 (2×10^{-4})	86.87 (1×10^{-4})	85.89 (4×10^{-5})	90.45 (5×10^{-5})	88.34 (8×10^{-4})	86.20 (6×10^{-4})
Spatial STAPLE	96.11 (0.13)	84.55 (2×10^{-4})	96.91 (0.54)	86.20 (0.004)	86.76 (0.003)	85.57 (0.009)	91.01 (2×10^{-4})	88.50 (5×10^{-4})	86.33 (9×10^{-4})
MaxMig	96.43 (0.13)	87.49 (0.001)	97.80 (0.53)	88.45 (0.006)	88.78 (0.005	89.70 (0.007)	92.85 (7×10^{-4})	90.21 (4×10^{-4})	87.62 (2×10^{-4})
Diag	96.75 (0.11)	91.22 (3×10^{-4})	97.93 (0.30)	90.79 (5×10^{-4})	–	–	–	–	–
DivP (ours)	**97.55**	**92.76**	**98.70**	**91.53**	**89.70**	**91.01**	**94.29**	**92.26**	**91.12**

$\mathscr{D}_{xx}$ represents the DICE coefficient of the result of the *xx* segmentation task. The *P* value, which is calculated using two-sample independent *t*-test, reflects the significance analysis of the results of MrPrism and those of other methods, with $P < 0.05$ indicating a significant difference in results. Bold indicates the best results, and '-' indicates that the test was not conducted on this dataset.

Table 4
Ablation study of the integration strategies and the loss functions on the brain tumor segmentation and OD/OC segmentation. Bold indicates the best results.

Integration			Loss			Brain tumor	OD/OC (REFUGE)	
Concat	ConvLSTM	ConvAtt	$\mathscr{L}_{MH}$	$\mathscr{L}_{ssf}$	$\mathscr{L}_{MH}+\mathscr{L}_{ssf}$	$\mathscr{D}_{\text{tumor}}$	$\mathscr{D}_{\text{disc}}$	$\mathscr{D}_{\text{cup}}$
✓				✓		86.14	94.45	85.38
	✓			✓		86.73	94.81	85.78
		✓	✓			86.69	94.27	84.89
		✓			✓	88.52	95.10	87.21
		✓		✓		**89.58**	**95.67**	**88.74**

$\mathscr{L}_{MH}$ and $\mathscr{L}_{ssf}$ indicate the multi-head loss and the proposed shuffle loss, respectively. Bold indicates the best results.

Table 5
Segmentation effect under Rec2 of MrPrism on different segmentation tasks of different datasets when different losses are used for Eqn. (8).

	OD/OC (REFUGE)		OD/OC (RIGA)		Brain growth	Brain tumor	Prostate1	Prostate2	Kidney
	$\mathscr{D}_{\text{disc}}$	$\mathscr{D}_{\text{cup}}$	$\mathscr{D}_{\text{disc}}$	$\mathscr{D}_{\text{cup}}$	$\mathscr{D}_{\text{brain}}$	$\mathscr{D}_{\text{tumor}}$	$\mathscr{D}_{\text{pros 1}}$	$\mathscr{D}_{\text{pros 2}}$	$\mathscr{D}_{\text{kidney}}$
MSE	94.21	88.03	95.28	87.76	84.73	88.01	86.85	76.10	76.27
SSIM	95.67	88.74	96.66	88.15	85.50	89.58	88.45	77.29	76.51

tions by taking advantage of iterative self-calibrated segmentation. It is noteworthy that in Table 1, for the OD segmentation task in the REFUGE dataset, the WDNet and MaxMig methods outperform our method by 0.04 in terms of DICE and for the OD segmentation task in the RIGA dataset, the Diag method outperforms the proposed method by 0.08 in terms of DICE. This is first due to the maturity of the task. From the significance analysis of the results of each method and the proposed methods in Table 1, it can be seen that all methods complete the OD segmentation task well, easily reaching the Dice of about 95%, and there is no significant difference between the results. Second, implementing the Diag method requires diagnostic task information, and among all the experiments in this paper, it can only be carried out on the optic cup and disc segmentation task of REGUGE and RIGA datasets. MrPrism effectively calibrates the segmentation results through recurrences, even in the condition of significant inter-rater variety. For example, on optic cup segmentation where the inter-rater variety is large, MrPrism gets 85.28% in term of DICE on RIGA dataset in the first recurrence and improves by 3.02% after four recurrences, achieving an 88.30% eventually. These experimental results demonstrate that MrPrism is able to calibrate the results by itself through the recurrences.

Fig. 6 illustrates the inter-rater variability between different experts for each task of various datasets. Specifically, for every pair of experts, their Dice values are computed based on the annotated segmentation labels. These values are stored in a matrix, and subsequently, the average Dice value is calculated for each expert concerning the labels of other experts. The label with the highest average Dice value is considered as the optimal segmentation label. As shown in the figure, OD segmentation task (Fig. 6a and c) has the smallest difference in multi-rater annotation, while OC segmentation task (Fig. 6b and d) has the largest, in consistency with what we found from our experimental results. In addition, the lung nodule segmentation task shown in Fig. 6j also exhibits inter-rater variability among the four experts' annotations. It can be observed that segmenting the optic cup in color fundus images and segmenting lung nodules in low-dose CT images are two highly challenging segmentation tasks, as targets are small and boundaries are uncertain. However, our experimental results (Table 1) validate that the proposed method in these challenging segmentation tasks still achieves SOTA performance.

To further explore the discrepancies between different methods, we compare the performance of the methods on multi-rater annotations respectively. The results are shown in Fig. 7. The experiment is conducted on two different segmentation tasks: Optic cup segmentation (OC-REFUGE), in which the inter-observer variety is large, and prostate segmentation (Prostate1), in which the inter-observer variety is relatively small. Several representative multi-rater learning strategies are selected for the comparison, including the three recurrences of MrPrism, CM, MaxMig, and MRNet. Compared with other methods, we can see CM, which implicitly learns the multi-rater confidences is inclined to believe few dominant raters, e.g., R1 and R2 in OC-REFUGE or R4 in Prostate l. However, without the dynamic calibration, the dominant raters are often incorrect, e.g., the credible rater in OC-REFUGE is actually R5 but not R1 or R2, which leads to its confident but incorrect predictions. MaxMig, which explicitly learns the multi-rater confidences, is prone to maintain a balance between the multiple raters. However, this strategy does not work well when the inter-rater variety is large, e.g., MaxMig gains only 84.45% on OC-REFUGE, while MrPrism gains 88.74%. MRNet requires the multi-rater confidences provided for the prediction. Under the default majority vote setting, its performance is close to the un-calibrated MrPrism-Rec0. Different from existing strategies, MrPrism is more adaptable to different tasks, On OC-REFUGE where few raters are much better than the others, MrPrism keeps increasing on some of the raters in the recurrences, e.g., R1, R2. R4, R5, while dropping on the other rater, e.g., R6 R7. It indicates that MrPrsim is able to discriminate the better raters from the others in the recurrences under the constraint of the raw image prior. On the other side, on Prostate1 task where most raters are equally correct, MrPrism raises the overall performance of most raters in the recurrences but is not overly biased like CM.

To discuss the impact of the number of annotation experts on the segmentation tasks in this study, we conducted the following experiments: observing the variation in the performance of the proposed method on segmenting lung nodules in the LIDCIDRI dataset and segmenting optic cup in the REFUGE dataset when the number of annotation experts increased from 2 to 4 and 7 for each task respectively. Both tasks have significant uncertainty. The experimental results are shown in Table 6, indicating that as the number of experts increases, the segmentation performance achieved by the proposed method improves. However, as shown in the second row of Table 6, when the number of experts is relatively large, further increasing the number results in diminishing returns in segmentation performance improvement. Therefore, considering the number of experts involved in the datasets used in this study, it is advisable to choose 6 experts when creating

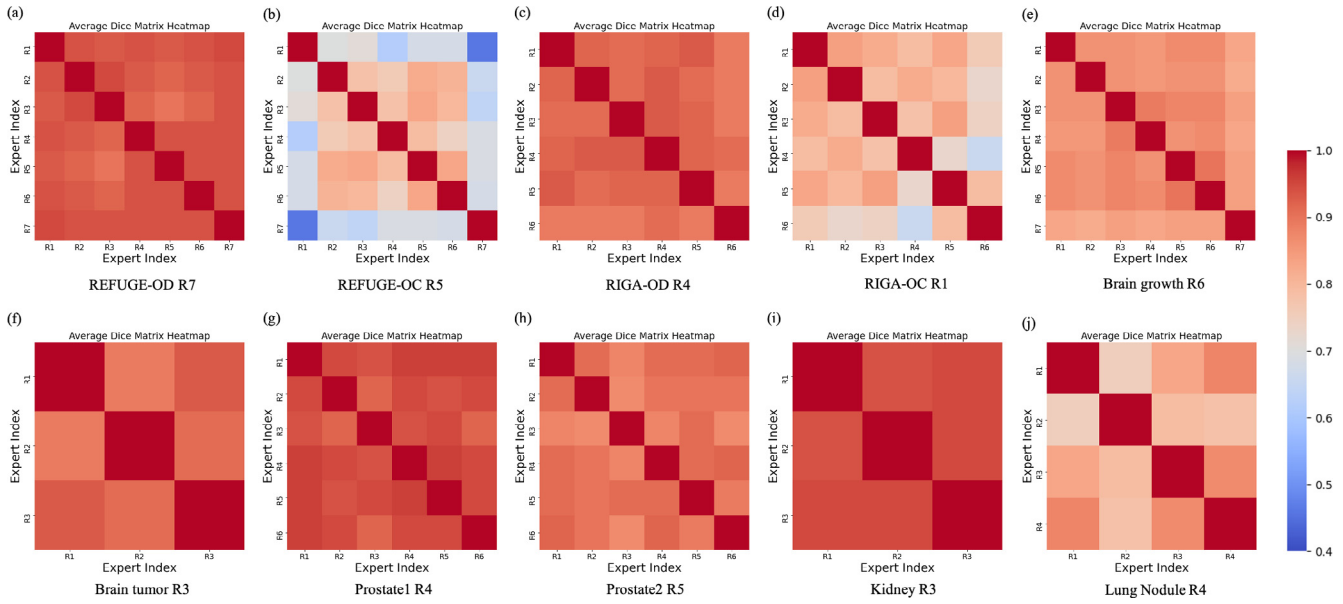


Fig. 6. The inter-rater variability in different datasets under different segmentation tasks. (a–j) These subgraphs are the dice matrixes between the segmentation labels annotated by each expert on different tasks of different datasets. The rater number corresponding to the best segmentation label is written in each matrix below. The degree of variability between experts can be observed from the colorbar; the closer to red, the smaller the variability.

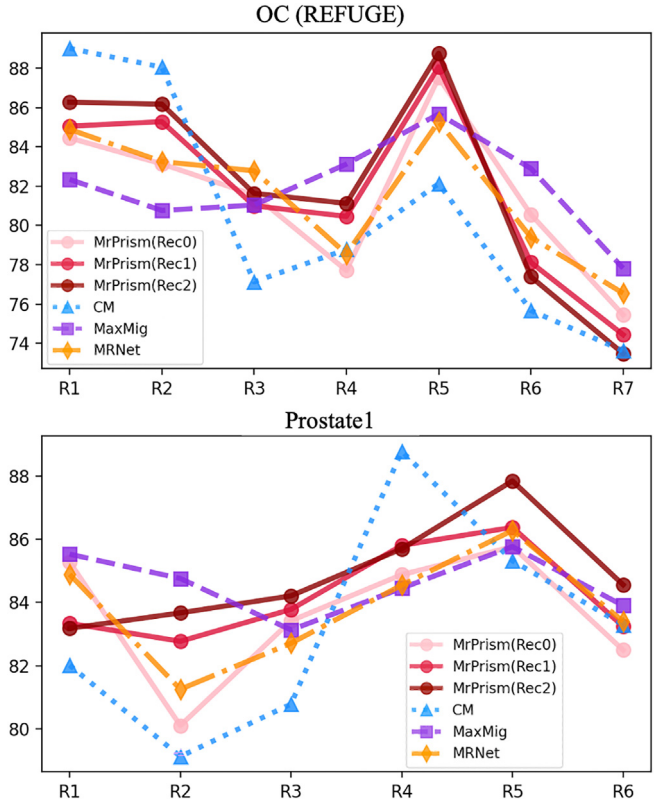


Fig. 7. The comparison of different methods on each of the multi-rater labels measured by dice coefficient (%). R *i* denotes the annotation of the rater *i*.

labels. However, in practical applications, determining the final number of experts should also take into account factors such as annotation difficulty and available manpower.

Although the converged segmentation effect of MrPrism has been significantly improved, it also has limitations, mainly in its anti-noise ability to the salt and pepper noise of the original image is limited. However, these noises can be removed by image quality enhancement algorithms.

In summary, we propose a self-calibrated segmentation model for learning medical segmentation from multi-rater labels. The model calibrates the segmentation and estimates the multi-rater confidences in a recurrent manner. By combining these two independent tasks, we can overcome their individual limitations and achieve mutual improvement through complementary strengths. Extensive empirical experiments demonstrate that our self-calibrated segmentation outperforms a wide range of alternative multi-rater learning methods.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

This work was supported by the Excellent Young Science and Technology Talent Cultivation Special Project of China Academy of Chinese Medical Sciences (CI2023D006), the National Natural Science Foundation of China (82121003 and 82022076), Beijing Natural Science Foundation (2190023), Shenzhen Fundamental Research Program (JCYJ20220818103207015), and Guangdong Provincial Key Laboratory of Human Digital Twin (2022B1212010004).

Table 6
Ablation study on the number of experts.

		R2	R3	R4	R5	R6	R7
LIDC-IDRI	$\mathscr{D}_{\text{nodule}}$	84.46	87.80	89.33	–	–	–
OC (REFUGE)	$\mathscr{D}_{\text{cup}}$	83.80	84.88	86.26	87.75	88.56	88.74

Author contributions

Junde Wu developed the concept of the manuscript and conducted the primary experiments. Huihui Fang, Jiayuan Zhu, Yu Zhang, and Xiang Li enriched the experimental design, conducted the result analysis, and finalized the manuscript. Yuanpei Liu, Huiying Liu, Yueming Jin, and Weimin Huang revised the manuscript writing. Qi Liu and Cen Chen provided guidance and support for accelerating the experiments. Yanfei Liu, Li Xiao, Weihua Yang, and Yue Liu comprehensively supervised and guided the clinical settings, medical data, and clinical applications of this study. Lixin Duan and Yanwu Xu comprehensively controlled the technology and provided instructive suggestions for the technical content of the manuscript.

Appendix A. Supplementary materials

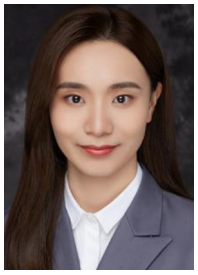
Supplementary materials to this article can be found online at https://doi.org/10.1016/j.scib.2024.06.037.

References

[1] Orlando JI, Fu HZ, Breda JB, et al. Refuge challenge: A unified frame-work for evaluating automated methods for glaucoma assessment from fundus photographs. Med Image Anal 2020;59:101570.
[2] Warrens MJ. Inequalities between multi-rater kappas. Adv Data Anal Classi 2010;4:271–86.
[3] Ji W, Yu S, Wu JD, et al. Learning calibrated medical image segmentation via multi-rater agreement modeling. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR). Vancouver; 2021; 12341–12351.
[4] Wu JD, Fang HH, Yang DL. et al. Opinions vary? diagnosis first! In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Singapore; 2022; 604-613.
[5] Dawid AP, Skene AM. Maximum likelihood estimation of observer error-rates using the EM algorithm. J R Stat Soc C-Appl 1979;28:20–8.
[6] Ghosh A, Kale S, McAfee P. Who moderates the moderators? crowdsourcing abuse detection in user-generated content. In: Proceedings of the 12th ACM conference on Electronic commerce (EC). San Jose; 2011; 167-176.
[7] N. Dalvi A. Dasgupta R. Kumar et al. Aggregating crowdsourced binary ratings In: Proceedings of the 22nd International Conference on World Wide Web (WWW). Rio de Janeiro; 2013; 285–294.
[8] Karger DR, Oh S, Shah D. Budget-optimal task allocation for reliable crowdsourcing systems. Oper Res 2014;62:1–24.
[9] Warfeld SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (staple): An algorithm for the validation of image segmentation. IEEE T Med Imaging 2004;23:903–21.
[10] Rodrigues F, Pereira F. Deep learning from crowds. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). New Orleans; 2018; 1611–1618.
[11] Albarqouni S, Baur C, Achilles F, et al. AggNet: Deep learning from crowds for mitosis detection in breast cancer histology images. IEEE T Med Imaging 2016;35:1313–21.
[12] Cao P, Xu YL, Kong YQ, et al. MaxMig: An information theoretic approach for joint learning from crowds. arXiv:1905.13436, 2019.
[13] Asman AJ, Landman BA. Robust statistical label fusion through consensus level, labeler accuracy, and truth estimation (COLLATE). IEEE T Med Imaging 2011;30:1779–94.
[14] Asman AJ, Landman BA. Formulating spatially varying performance in the statistical fusion framework. IEEE T Med Imaging 2012;31:1326–36.
[15] Tanno R, Saeedi A, Sankaranarayanan S, et al. Learning from noisy labels by regularized estimation of annotator confusion. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR). Long Beach; 2019; 11244–11253.
[16] Guan MY, Gulshan V, Dai AM, et al. Who said what: Modeling individual labelers improves classification. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). New Orleans; 2018; 2668–3603.
[17] Chou HC, Lee CC. Every rating matters: Joint learning of subjective labels and individual annotators for speech emotion classification. In: Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton; 2019; 5886–5890.
[18] Rupprecht C, Laina I, DiPietro R, et al. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In: Processing of the International Conference on Computer Vision (ICCV). Venice; 2017; 3591–3600.
[19] Jensen MH, Jørgensen DR, Jalaboi R, et al. Improving uncertainty estimation in convolutional neural networks using inter-rater agreement. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Shenzhen; 2019; 540–548.
[20] Kendall A, Badrinarayanan V, Cipolla R. Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. In: Proceeding of the British Machine Vision Conference (BMVC). London; 2017; 57.1–57.12.
[21] Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? Adv Neural Inform Process Syst 2017;30:5574–84.
[22] Lee S, Purushwalkam S, Cogswell M, et al. Stochastic multiple choice learning for training diverse deep ensembles. Adv Neural Inform Process Syst 2016;29:2119–27.
[23] Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. Adv Neural Inform Process Syst 2017;30:6402–13.
[24] Baumgartner CF, Tezcan KC, Chaitanya K, et al. PHiSeg: Capturing uncertainty in medical image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Shenzhen; 2019 119–127.
[25] Kohl SAA, Romera-Paredes B, Meyer C, et al. A probabilistic U-Net for segmentation of ambiguous images. Adv Neural Inform Process Syst 2018;31:6965–75.
[26] Kohl SAA, Romera-Paredes B, Maier-Hein KH, et al. A hierarchical probabilistic U-Net for modeling multi-scale ambiguities. arXiv:1905.13077, 2019.
[27] Jungo A, Meier R, Ermis E, et al. On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Granada; 2018; 682–690.
[28] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Adv Neural Inf Process Syst 2017;30:5998–6008.
[29] Almazroa A, Alodhayb S, Osman E, et al. Agreement among ophthalmologists in marking the optic disc and optic cup in fundus images. Int Ophthalmol 2017;37:701–17.
[30] Mehta R, Filos A, Baid U, et al. QU-BraTS: MICCAI BraTS 2020 challenge on quantifying uncertainty in brain tumor segmentation-analysis of ranking scores and benchmarking results. J Mach Learn Biomed Imaging 2022: 2022.
[31] Menze B, Joskowicz L, Berger C, et al. Quantification of uncertainties in biomedical image quantification. Zenodo 2020.
[32] Armato III SG, McLennan G, Bidaut L, et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans. Med Phys 2011;38:915–31.
[33] Geman D, Reynolds G. Constrained restoration and the recovery of discontinuities. IEEE Trans Pattern Anal Mach Intell 1992;14:367–83.
[34] Zhang K, Zuo WM, Gu SH, et al. Learning deep CNN denoiser prior for image restoration. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR). Hawaii; 2017; 3929-3938.
[35] Dong WS, Wang PY, Yin WT, et al. Denoising prior driven deep neural network for image restoration. IEEE Trans Pattern Anal Mach Intell 2018;41:2305–18.
[36] Wu JD, Di XG, et al. Integrating neural networks into the blind deblurring framework to compete with the end-to-end learning-based methods. IEEE Trans Image Process 2020;29:6841–51.
[37] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers. In: Proceeding of the European Conference on Computer Vision (ECCV). Glasgow; 2020; 213–229.
[38] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. arXiv:2010.11929, 2020.
[39] He KM, Zhang XY, Ren SQ, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas; 2016; 770–778.
[40] Ba JL, Kiros JR, Hinton GE. Layer normalization. arXiv:1607.06450, 2016.
[41] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Munich; 2015; 234–241.
[42] Chollet F. Xception: Deep learning with depth-wise separable convolutions. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR). Hawaii; 2017; 1251–1258.
[43] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv:1412.6980, 2014.
[44] Milletari F, Navab N, Ahmadi SA. V-Net: Fully convolutional neural net-works for volumetric medical image segmentation. In: The 4th international conference on 3D Vision (3DV). Palo Alto; 2016; 565–571.
[45] Karimi D, Gholipour A. Improving calibration and out-of-distribution detection in deep models for medical image segmentation. IEEE Trans Artif Intell 2022;4:383–97.
[46] Yeung M, Rundo L, Nan Y, et al. Calibrating the dice loss to handle neural network overconfidence for biomedical image segmentation. J Digit Imaging 2023;36:739–52.
[47] Rahman A, Valanarasu JMJ, Hacihaliloglu I, et al. Ambiguous medical image segmentation using diffusion models. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR). Vancouver; 2023; 11536–11546.
[48] Zhang SH, Fu HZ, Yan YG, et al. Attention guided network for retinal image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Shenzhen; 2019; 797–805.

[49] Isensee F, Jaeger PF, Kohl SA, et al. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18:203–11.
[50] Chen JN, Lu YY, Yu QH, et al. TransUNet: Transformers make strong encoders for medical image segmentation. arXiv: 2102.04306, 2021.
[51] Shi XJ, Chen ZR, Wang H, et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Adv Neural Inform Process Syst 2015;28:802–10.

Junde Wu received his B.S. and M.D. degrees from Harbin Institute of Technology. He is now a Ph.D. candidate at Oxford University. He is also a research associate at the National University of Singapore. His research interest focuses on foundational models in medicine, multi-rater labels fusion, and diffusion models.



Huihui Fang received her B.S. and Ph.D. degrees from the Beijing Institute of Technology. She is currently working as an associate researcher in Pazhou Laboratory. Her research interests include image processing in ophthalmology, application of deep learning models in medical settings, and foundational models.

Yanwu Xu is a professor at South China University of Technology and a researcher at Pazhou Laboratory. He is also an expert in WHO Digital Health Advisory Committee, and a visiting professor at Singapore Eye Institute. His research interest includes pattern recognition, intelligent healthcare, active healthcare management, and artificial intelligence.

Li Xiao, Ph.D., an associate professor and Master's supervisor of University of Electronic Science and Technology of China, an associate chief physician of Stomatology of Sichuan Provincial People's Hospital. His research focuses on osteoarthritis, especially the molecular mechanism and targeted therapy of osteoarthritis.

Weihua Yang, Ph.D., is the Chief Physician of Ophthalmology, the Director of the Office of Big Data and Artificial Intelligence at Shenzhen Eye Hospital, and the Executive Deputy Director of the Shenzhen Eye Institute. His research focuses on basic and clinical ophthalmology.

Yue Liu, Ph.D., is currently a full professor at Cardiovascular Disease Center, Xiyuan Hospital of China Academy of Chinese Medical Sciences. His research focuses on the basic and clinical cardiology of integrative medicine.