Learning Calibrated Medical Image Segmentation via Multi-rater Agreement Modeling
通過多評估者協議建模學習校準的醫學圖像分割

Wei Ji, Shuang Yu, Junde Wu, Kai Ma, Cheng Bian, Qi Bi, Jingjing Li, Hanruo Liu, Li Cheng, Yefeng Zheng

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

CVPR 2021 Best Paper Candidate
Paper(Best Paper Candidate):
https://openaccess.thecvf.com/content/CVPR2021/html/Ji_Learning_Calibrated_Medical_Image_Segmentation_via_Multi-Rater_Agreement_Modeling_CVPR_2021_paper.html
Code: https://github.com/jiwei0921/MRNet/

Abstract 摘要

In medical image analysis, it is typical to collect multiple annotations, each from a different clinical expert or rater, in the expectation that possible diagnostic errors could be mitigated.
在醫學圖像分析中，通常會收集多個註釋，每個註釋來自不同的臨床專家或評估者，以期減少可能的診斷錯誤。

Meanwhile, from the computer vision practi-tioner viewpoint, it has been a common practice to adopt the ground-truth labels obtained via either the majority-vote or simply one annotation from a preferred rater.
同時，從計算機視覺從業者的角度來看，採用通過多數票或來自首選評估者的簡單註釋獲得的真實標籤已成為一種常見做法。

This process, however, tends to overlook the rich information of agreement or disagreement ingrained in the raw multi-rater annotations.
然而，這個過程往往會忽略在原始多評估者註釋中根深蒂固的同意或不同意的豐富資訊。

To address this issue, we propose to ex-plicitly model the multi-rater (dis-)agreement, dubbed MR-Net, which has two main contributions.
為了解決這個問題，我們建議對多評價者（不）協議進行顯式建模，稱為 MR-Net，它有兩個主要貢獻。

First, an expertise-aware inferring module or EIM is devised to embed the expertise level of individual raters as

prior knowledge, to form high-level semantic features.
首先，設計了一個專業知識推斷模塊或 EIM，將各個評分者的專業知識水平作為先驗知識嵌入，以形成高級語義特徵。

Second, our approach is capable of reconstructing multi-rater gradings from coarse predictions, with the multi-rater (dis-)agreement cues being further exploited to improve the segmentation performance.
其次，我們的方法能夠從粗略預測中重建多評分者的評分，並進一步利用多評分者（不一致）線索來提高分割性能。

To our knowledge, our work is the first in producing cali-brated predictions under different expertise levels for medical image segmentation.
據我們所知，我們的工作是第一個在不同專業水平下為醫學圖像分割生成校準預測的工作。

Extensive empirical experiments are conducted across five medical segmentation tasks of diverse imaging modalities.
在不同成像方式的五個醫學分割任務中進行了廣泛的實證實驗。

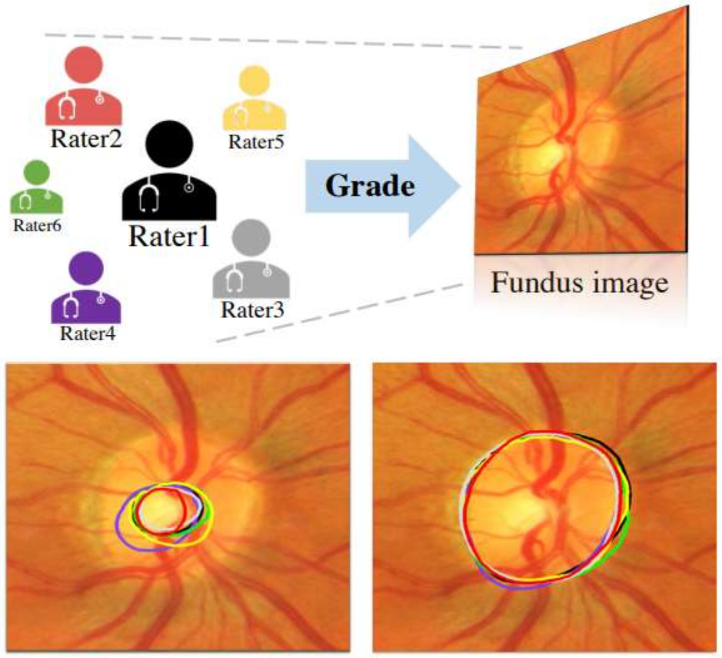
In these experiments, superior performance of our MRNet is observed comparing to the state-of-the-arts, indicating the effectiveness and applicability of our MRNet toward a wide range of medical segmentation tasks.
在這些實驗中，與現有技術相比，我們觀察到了我們的 MRNet 的優越性能，表明我們的 MRNet 對廣泛的醫學分割任務的有效性和適用性。

Source code is publicly available.

1. Introduction 前言

Accurate anatomy and lesion segmentation is crucial in clinical assessment of various diseases, including for example glaucoma [28, 36, 43], prostate diseases [30, 52], and brain tumors [11, 17, 44].
準確的解剖結構和病變分割對於各種疾病的臨床評估至關重要，包括青光眼 [28、36、43]、前列腺疾病 [30、52] 和腦腫瘤 [11、17、44]。

It has been increasingly popular to develop automated segmentation systems, to facilitate a reliable reference for the quantification of disease progression, which is especially accelerated by the exciting breakthroughs of deep convolutional neural networks (CNNs) [7, 20, 34, 35, 49, 55, 56, 59] over the past decade.
開發自動分割系統越來越受歡迎，以促進量化疾病進展的可靠參考，尤其是深度卷積神經網絡 (CNN) [7, 20, 34, 35, 49, 55, 56, 59] 在過去十年中。



Optic cup annotations　Optic disc annotations

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.

Figure 1. Top: an exemplar medical image grading scenario conducted by multiple raters with different expertise levels.
圖 1. 頂部：由具有不同專業水平的多個評估者進行的示例性醫學圖像分級場景。

Bottom: visualization of optic cup and disc annotations of the above raters.
底部：上述評估者的視杯和視盤註釋的可視化。

Different from labelling natural images, medical images are often independently annotated by a group of experts or raters, to mitigate the subjective bias of a particular rater due to factors such as the level of expertise, or possible negligence of subtle symptoms [13, 39, 23, 28].
與標記自然圖像不同，醫學圖像通常由一組專家或評估者獨立註釋，以減輕特定評估者由於專業水平或可能疏忽細微症狀等因素而產生的主觀偏見 [13, 39, 23, 28]。

Inter-observer variability, as frequently reported by relevant research in the clinical field, often leads to challenges in segmenting highly uncertain regions [3, 23, 37].
正如臨床領域相關研究經常報告的那樣，觀察者間的變異性通常會導致分割高度不確定區域的挑戰 [3, 23, 37]。

Fig. 1 provides a representative illustration of the multi-rater grading process in annotating optic cups and discs from fundus images, with notable uncertainties or disputed regions presented among graders.
圖 1 提供了從眼底圖像註釋視杯和視盤的多評級者分級過程的代表性說明，在分級者之間存在明顯的不確定性或有爭議的區域。

It is thus necessary for automated systems to consider a proper segmentation strategy that reflects the underlying (dis-)agreement among multiple experts.
因此，自動化系統有必要考慮一種適當的分割策略，以反映多個專家之間的潛在（不一致）意見。

Existing works typically require unique ground-truth annotations, each pairing with one of the input images to train the deep learning models.
現有的工作通常需要獨特的地面實況註釋，每個註釋與輸入圖像之一配對以訓練深度學習模型。

It is a common practice to take majority vote, STAPLE [50] or other label fusion strategies to obtain the ground-truth labels [5, 29, 30, 34, 57, 59].
採用多數投票、STAPLE [50] 或其他標籤融合策略來獲得真實標籤 [5, 29, 30, 34, 57, 59] 是一種常見的做法。

Being simple and easy to implement, this strategy, however, comes at the cost of ignoring altogether the underlying uncertainty information among multiple experts.
然而，這種策略簡單且易於實施，其代價是完全忽略了多個專家之間的潛在不確定性資訊。

Very recently, several efforts start to explore the influence of multi-rater labels by label sampling [19, 24] or multi-head [16] strategies.
最近，一些努力開始通過標籤採樣 [19, 24] 或多頭 [16] 策略探索多評價者標籤的影響。

It is reported that models trained with multi-rater labels are better calibrated than those with the typical ground-truth label via, e.g. majority vote, which are prone to be overconfident [19, 24].
據報導，使用多評分者標籤訓練的模型比使用典型地面實況標籤訓練的模型更好地校準，例如 多數票，這容易過度自信 [19, 24]。

Meanwhile, there still lacks a principled approach to incorporate in training the rich uncertainty information from multiple raters.
同時，仍然缺乏一種原則性的方法來將來自多個評估者的豐富的不確定性資訊納入訓練。

Specifically, we focus on the following questions:
具體來說，我們關注以下問題：

1) how to integrate varied expertise-level, or expertness, of individual raters into the network architecture?
1）如何將各個評估者的不同專業水平或專業知識整合到網絡架構中？

2) how to exploit the uncertainty information among different experts to produce probability maps that better reflect the underlying graders' (dis-)agreement?
2）如何利用不同專家之間的不確定性信息來生成更好地反映潛在評分者（不）同意的概率圖？

This inspires us to propose a multi-rater agreement modeling framework, MRNet.
這激勵我們提出一個多評估者協議建模框架，MRNet。

To our knowledge, it is the first in explicitly addressing the above-mentioned questions.
據我們所知，它是第一個明確解決上述問題的。

Our framework has the following three main contributions:
我們的框架有以下三個主要貢獻：

・ The notion of expertness is explicitly introduced as prior knowledge about the expertise levels of the involved multi-raters.
・ 專業度的概念被明確引入，作為有關所涉及的多評估者專業度水平的先驗知識。

It is embedded in the high-level semantic features through the proposed Expertiseaware Inferring Module (EIM), enabling the representation capability to accommodate the multi-rater settings.
它通過提議的專業知識推斷模塊 (EIM) 嵌入到高級語義特徵中，使表示能力能夠適應多評估者設置。

・ A Multi-rater Reconstruction Module (MRM) is designed to reconstruct the raw multi-rater gradings from the the expertness prior and the soft prediction of the model.
・ 多評分者重建模塊 (MRM) 旨在根據專家先驗和模型的軟預測重建原始多評分者評分。

This enables the estimation of an uncertainty map that reflects the inter-rater variability, by exploiting the intrinsic correlations between the fused soft label and the raw multi-rater annotations.
通過利用融合軟標籤和原始多評級者註釋之間的內在相關性，這使得能夠估計反映評級者間可變性的不確定性圖。

・ To better utilize the rich cues among multi-rater (dis-)agreements, we further incorporate in our framework a Multi-rater Perception Module (MPM), which empirically leads to noticeable performance boost.
・ 為了更好地利用多評價者（不）協議之間的豐富線索，我們進一步在我們的框架中加入了多評價者感知模塊（MPM），這在經驗上導致顯著的性能提升。

Extensive experiments are performed on five different medical image segmentation tasks of diverse image modalities, including color fundus imaging, computed tomography (CT), and magnetic resonance imaging (MRI).
・ 為了更好地利用多評價者（協議不之間）的文字，我們進一步在我們的框架中加入了多評價者模塊（MPM），這在經驗上導致顯著的業績提升。

Overall, our MRNet framework consistently outperforms the stateof-the-art methods as well as existing multi-rater strategies.
總體而言，我們的 MRNet 框架始終優於最先進的方法以及現有的多評估者策略。

In addition, our MRNet runs in real-time (29 frame per second) at inference stage, making it practically appealing for many real-world applications.
此外，我們的 MRNet 在推理階段實時運行（每秒 29 幀），使其對許多實際應用程序具有實際吸引力。

2. Related Work 相關工作

Medical Image Segmentation. 醫學圖像分割。

With the advancement of CNNs, an increasing number of deep learning architectures have been proposed for medical segmentation tasks such as optic disc/cup segmentation [60, 29, 57, 12] in fundus images, prostate segmentation [21, 30, 48] and brain tumor segmentation [4, 6].
隨著 CNN 的進步，越來越多的深度學習架構被提出用於醫學分割任務，例如眼底圖像中的視盤/杯分割 [60、29、57、12]、前列腺分割 [21、30、48] 和 腦腫瘤分割 [4, 6]。

These methods have obtained superior performance comparing to traditional feature engineering based methods [8, 9, 10].
與傳統的基於特徵工程的方法相比，這些方法獲得了優越的性能 [8, 9, 10]。

Taking optic disc/cup segmentation as an example, Fu et al. [12] proposed a U-shaped network with multi-scale supervision strategy for polar transformed fundus images to produce the segmentation maps.
以視盤/杯分割為例，Fu et al. [12]提出了一種具有多尺度監督策略的 U 形網絡，用於極地變換眼底圖像以生成分割圖。

Gu et al. [15] integrated dense atrous convolution block and residual multi-kernel pooling to U-Net structure to capture high-level features with context information.
Gu et al. [15] 將密集的多孔卷積塊和殘差多核池化集成到 U-Net 結構中，以捕獲具有上下文信息的高級特徵。

Zhang et al. [58] presented an attention guided network using guided filter to preserve the structural information and reduce the negative influence of background.
Zhang et al. [58] 提出了一個注意力引導網絡，使用引導過濾器來保留結構信息並減少背景的負面影響。

Meanwhile, Li et al. [29] integrated detection and multi-class segmentation into a unified architecture for segmenting the optic cup and disc regions.
與此同時，Li et al. [29] 將檢測和多類分割集成到一個統一的架構中，用於分割視杯和視盤區域。

Wang et al. [45] attempted to utilize the designed domain adaptation frameworks for fundus image segmentation, in order to increase the cross-domain prediction accuracy.
Wang et al. [45]嘗試利用設計的域自適應框架進行眼底圖像分割，以提高跨域預測精度。

A common practice adopted by the above-mentioned methods, as well as most existing CNNs based learning methods, is to construct training examples by retaining unique ground-truth labels for each of the training instances.
上述方法以及大多數現有的基於 CNN 的學習方法採用的常見做法是通過為每個訓練實例保留唯一的真實標籤來構建訓練實例。

In this manner, the valuable multi-rater labels obtained in the grading procedure with inter-rater variability are unfortunately not well-exploited.
以這種方式，不幸的是，在評分過程中獲得的有價值的多評分者標籤具有評分者間可變性，但沒有得到很好的利用。

Multi-rater Strategies. 多評級策略。

Very recently, the problems of the multi-rater labels and inter-rater variability start to attract research attentions [16, 19, 2, 24, 42, 54]. Jensen et al. [19] adopted a label sampling strategy for skin disease classification, by sampling labels randomly from the multi-rater labeling pool during each training iteration.
最近，多評分者標籤和評分者間可變性的問題開始引起研究關注 [16, 19, 2, 24, 42, 54]。 詹森等人。 [19] 採用標籤採樣策略進行皮膚病分類，在每次訓練迭代期間從多評價者標籤池中隨機採樣標籤。

It was observed that model trained with the traditional unique ground-truths would be over-confident, meanwhile model trained with label sampling strategy was better calibrated.
觀察到，使用傳統的獨特地面實況訓練的模型會過度自信，同時使用標籤採樣策略訓練的模型可以更好地校準。

Similar observation was also reported by [24] for segmentation task.
[24] 對分割任務也報告了類似的觀察結果。

Label sampling strategy was also utilized by [26] to train a probabilistic model based on a combined U-Net with conditional variational autoencoder to obtain multiple plausible hypotheses.
[26] 還利用標籤採樣策略來訓練基於組合 U-Net 和條件變分自動編碼器的概率模型，以獲得多個似是而非的假設。

Similarly, Baumgartner et al. [2] employed label sampling strategy as well to train the hierarchical probabilistic model with multi-scale latent variables when using labels from multiple annotators.
類似地，Baumgartner et al. [2] 在使用來自多個註釋器的標籤時，也採用標籤採樣策略來訓練具有多尺度潛在變量的分層概率模型。

Guan et al. [16] predicted the gradings of each rater individually and learned the corresponding weights for final prediction.
Guan et al. [16] 分別預測了每個評分者的評分，並學習了相應的權重以進行最終預測。

Yu et al. [54] proposed a multi-branch structure to generate three predictions under different sensitivity settings, to leverage multi-rater consensus information for glaucoma classification.
Yu et al. [54] 提出了一種多分支結構，以在不同的靈敏度設置下生成三個預測，以利用多評估者共識信息進行青光眼分類。

With the existing multi-rater strategies of label sampling [19, 2, 24] and multiple-head/branch architecture [16, 54], there still lacks a principled research investigation on exploiting the rich (dis-)agreement information among raters in model training and predictions.
使用現有的標籤抽樣 [19, 2, 24] 和多頭/分支架構 [16, 54] 的多評估者策略，仍然缺乏關於利用評估者之間豐富的（不一致）信息的原則性研究調查 模型訓練和預測。

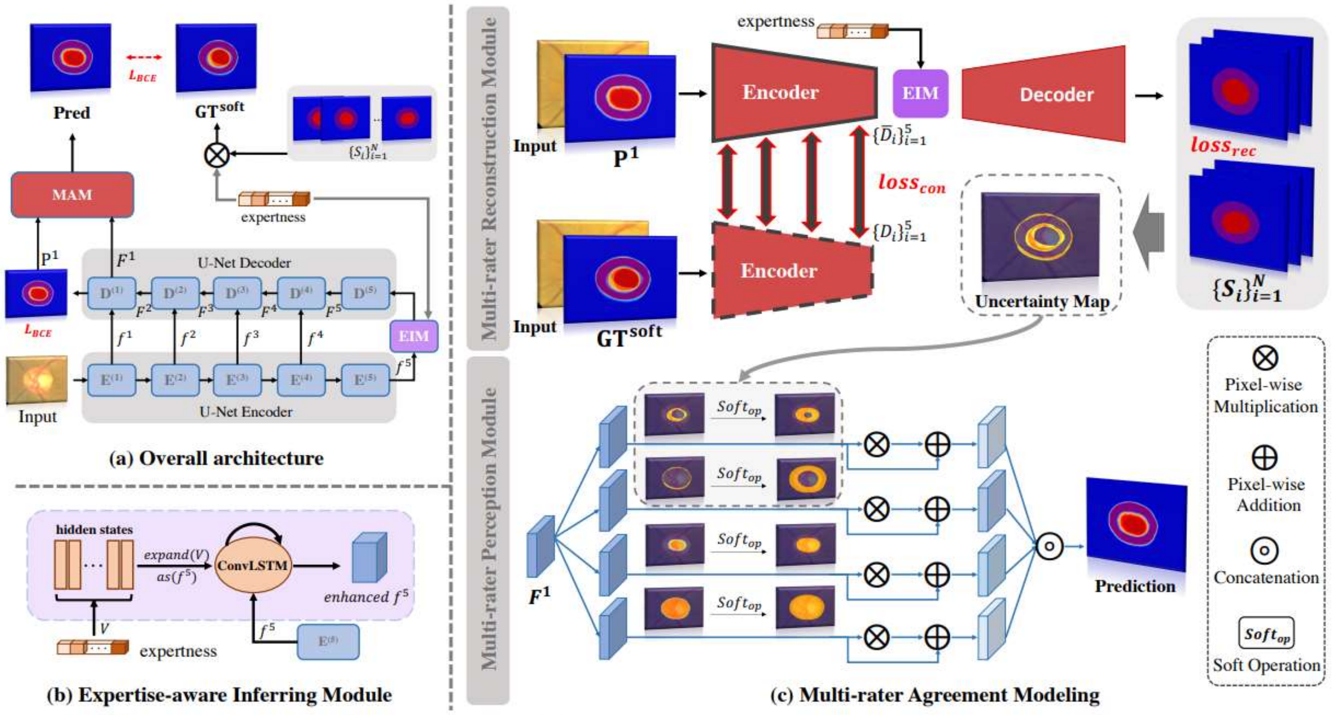


Figure 2. An illustration of our MRNet framework, which starts from (a) an overview of the processing pipeline, and continues with zoomed-in diagrams of individual modules, including (b) the Expertise-aware Inferring Module (EIM), and (c) the Multi-rater Agreement Modeling (MAM) that consists of the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

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).
圖 2. 我們的 MRNet 框架的圖示，從 (a) 處理管道的概述開始，然後是各個模塊的放大圖，包括 (b) 專業知識推斷模塊 (EIM)，以及 (c) 多評價者協議建模 (MAM)，由多評價者重建模塊 (MRM) 和多評價者感知模塊 (MPM) 組成。

3. Methodology 方法

3.1. Motivation 動機

As aforementioned, the inter-grader variability is a wellknown issue in the medical image annotation process, since experts differ from each other in their grading preferences and levels of expertise [13, 39, 23, 28].
如前所述，分級者間的可變性是醫學圖像註釋過程中的一個眾所周知的問題，因為專家在分級偏好和專業水平方面彼此不同 [13, 39, 23, 28]。

In order to quantitatively demonstrate such difference, a preliminary experiment is performed with an optic cup segmentation setting on the RIGA benchmark dataset [1].
為了定量證明這種差異，在 RIGA 基準數據集 [1] 上使用視杯分割設置進行了初步實驗。

We train a U-Net [38] using individual rater's annotations for the optic cup segmentation task, and thus obtain six different models (named Model 1-6) corresponding to six raters (named Rater 1-6).
我們針對視杯分割任務使用單個評分者的註釋訓練 U-Net [38]，從而獲得對應於六個評分者（稱為評分者 1-6）的六個不同模型（稱為模型 1-6）。

The performance of each model against each rater's grading as well as the final consensus label from majority vote is listed in Table 1.
表 1 列出了每個模型針對每個評估者的評分以及多數投票的最終共識標籤的表現。

It is obvious that all the models have the optimal performance when trained and evaluated with the same rater's annotations but much worse when evaluated by others' annotations.
很明顯，所有模型在使用相同評分者的註釋進行訓練和評估時都具有最佳性能，但在使用其他人的註釋進行評估時要差得多。

Moreover, when evaluated with the consensus labels obtained with majority vote, Model 1 achieves the best result, followed by Model 2, which is consistent with the database and grader analysis reported in [1].
此外，當使用多數投票獲得的共識標籤進行評估時，模型 1 取得了最好的結果，其次是模型 2，這與 [1] 中報告的數據庫和分級器分析一致。

Two findings can therefore be drawn from here and possibly generalized to medical analysis tasks beyond optic cup segmentation:
因此，可以從這裡得出兩個發現，並可能推廣到視杯分割以外的醫學分析任務：

1) individual expert has specific and consistent grading patterns and
1) 個別專家具有特定且一致的評分模式，並且

2) the expertise levels among a group of graders are usually different from one to the other.
2) 一組評分者之間的專業水平通常各不相同。

This preliminary study and subsequent findings motivate us to propose our MRNet framework to be discussed next.

這項初步研究和隨後的發現促使我們提出我們接下來要討論的 MRNet 框架。

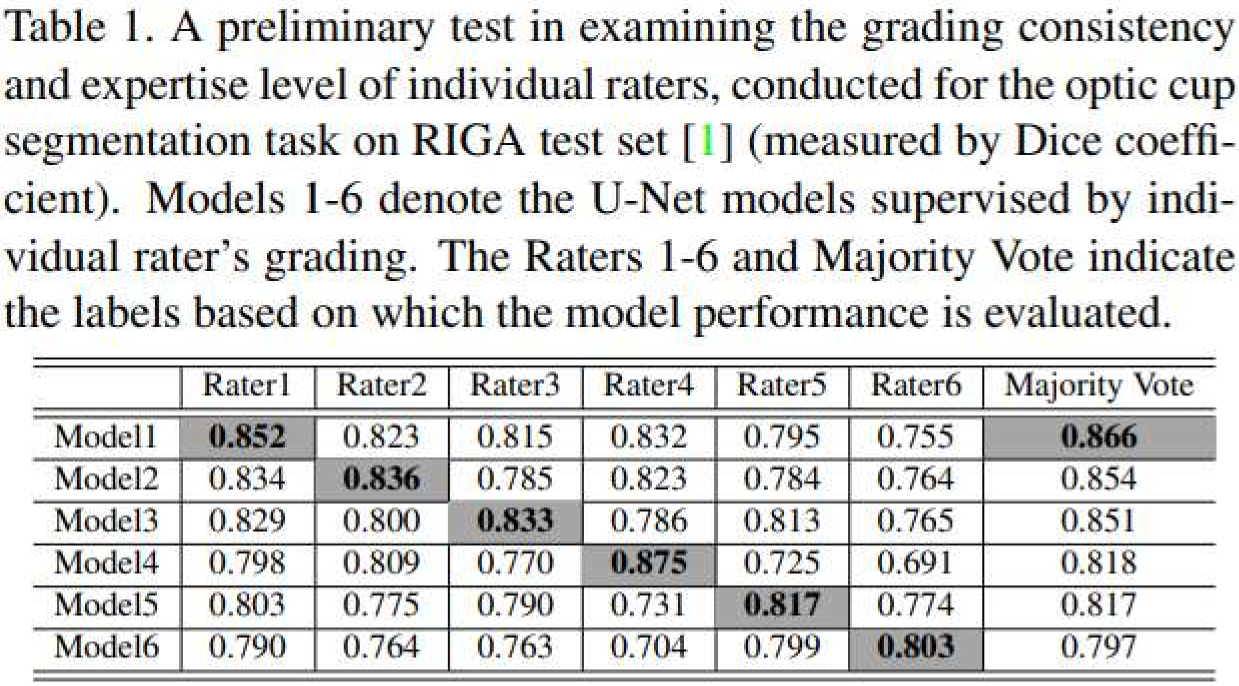


Table 1. A preliminary test in examining the grading consistency and expertise level of individual raters, conducted for the optic cup segmentation task on RIGA test set [1] (measured by Dice coefficient). Models 1-6 denote the U-Net models supervised by individual rater's grading. The Raters 1-6 and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

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).
表 1. 在 RIGA 測試集 [1]（通過 Dice 係數測量）上為視杯分割任務進行的檢查各個評分者的評分一致性和專業水平的初步測試。

Models 1-6 denote the U-Net models supervised by individual rater's grading.
模型 1-6 表示由個人評分者評分監督的 U-Net 模型。

The Raters 1-6 and Majority Vote indicate the labels based on which the model performance is evaluated.
評分者 1-6 和多數票表示評估模型性能所依據的標籤。

3.2. Overall Framework 總體框架

In this work, we propose a novel medical image segmentation framework, named as MRNet, that takes underlying agreement/disagreement information among multiple raters into consideration.
在這項工作中，我們提出了一種新的醫學圖像分割框架，稱為 MRNet，它考慮了多個評估者之間的潛在同意/不同意信息。

Fig. 2 illustrates the overall framework of the proposed MRNet, which contains a coarse to fine two-stage processing pipeline.
圖 2 說明了所提出的 MRNet 的整體框架，其中包含一個粗到細的兩級處理流水線。

The first stage adopts the widely used U-Net architecture [38] with a ResNet34 [18] backbone pretrained from

ImageNet as the encoder part.
第一階段採用廣泛使用的 U-Net 架構 [38]，並使用從 ImageNet 預訓練的 ResNet34 [18] 主幹作為編碼器部分。

Then an Expertise-aware Inferring Module (EIM) is inserted at the bottleneck layer to embed the expertise information of individual raters, named as expertness vector, into the extracted high-level semantic features of the network.
然後在瓶頸層插入一個專業知識推理模塊（EIM），將各個評分者的專業知識信息（稱為專業知識向量）嵌入到提取的網絡高級語義特徵中。

The enhanced feature f5 is further passed to the decoder blocks of U-Net to generate multi-level decoder features {Fi}5i =1.
增強特徵 f5 進一步傳遞給 U-Net 的解碼器塊以生成多級解碼器特徵 {Fi}5i =1。

The final decoded feature F1 is processed by a 1 × 1 convolutional operation followed by a sigmoid activation function to generate the coarse prediction P1.
最終解碼的特徵 F1 由 1 × 1 卷積運算處理，然後是 sigmoid 激活函數以生成粗預測 P1。

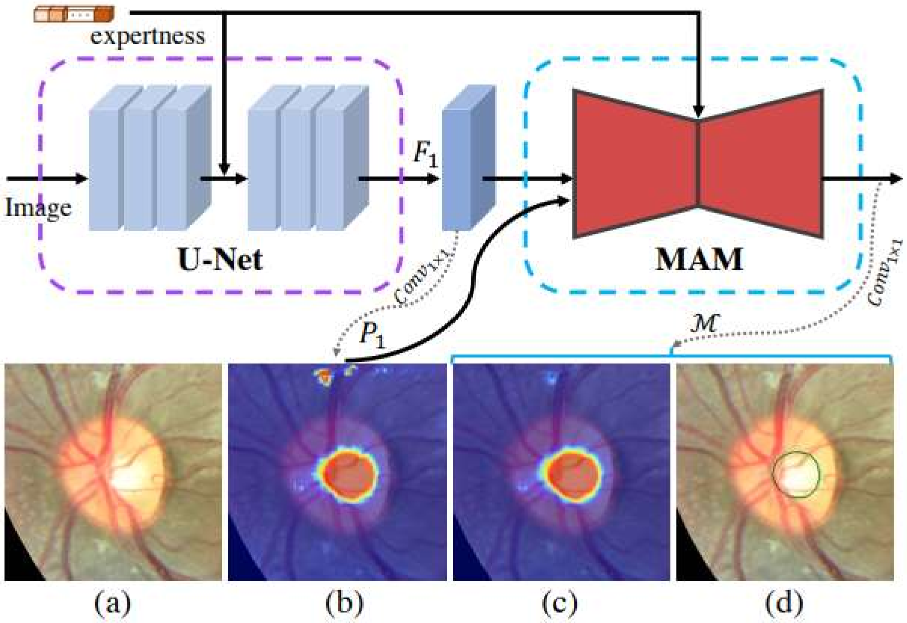


Figure 3. Intermediate visual results in the processing pipeline of our MRNet framework. (a) Input fundus image. (b) Heat map of the initial cup prediction P^1. (c) Heat map of the final refined cup prediction $\mathcal{M}$. (d) Segmentation boundary of the cup prediction (green) and ground-truth (black).

Figure 3. Intermediate visual results in the processing pipeline of our MRNet framework.
圖 3. MRNet 框架處理管道中的中間視覺結果。

(a) Input fundus image.
(a) 輸入眼底圖像。

(b) Heat map of the initial cup prediction P1.
(b) 初始杯預測 P1 的熱圖。

(c) Heat map of the final refined cup prediction M. (d) Segmentation boundary of the cup prediction (green) and ground-truth (black).
(c) 最終精煉杯預測 M 的熱圖。 (d) 杯預測（綠色）和地面實況（黑色）的分割邊界。

The second stage, aiming to refine the coarse prediction results from the first stage to get better predictions, is composed of two modules arranged in a sequential order.
第二階段旨在細化第一階段的粗略預測結果以獲得更好的預測，由按順序排列的兩個模塊組成。

The Multi-rater Reconstruction Module (MRM) is designed to reconstruct the raw multi-rater's gradings, based on which to estimate the pixel-wise uncertainty map that represents the inter-observer variability across different regions.
多評分者重建模塊 (MRM) 旨在重建原始多評分者的評分，在此基礎上估計代表不同區域的觀察者間變異性的像素級不確定性圖。

Furthermore, the Multi-rater Perception Module (MPM) with soft attention mechanism is proposed to utilize the uncertainty map to refine the coarse prediction.
此外，提出了具有軟注意力機制的多評級者感知模塊（MPM），以利用不確定性圖來細化粗略預測。

For simplicity, we use Multi-rater Agreement Modeling (MAM) to represent the combination of the two sequential modules. A simplified illustration of the pipeline with intermediate results is also shown in Fig. 3.
為簡單起見，我們使用多評級者協議建模 (MAM) 來表示兩個順序模塊的組合。 圖 3 還顯示了具有中間結果的流水線的簡化說明。

3.3. Expertiseaware 專業知識

Inferring Module 推理模塊

Considering that different experts have different levels of clinical expertise and thus should be assigned with different weights during the model training procedure, we propose an Expertise-aware Inferring Module (EIM) to take advantage of the expertise levels of individual raters as prior knowledge, which is embedded into the segmentation network in the format of conditional information to increase the dynamical representation capability of the extracted features.
考慮到不同專家具有不同水平的臨床專業知識，因此在模型訓練過程中應分配不同的權重，我們提出了專業知識推理模塊（EIM），以利用各個評估者的專業知識水平作為先驗知識， 以條件信息的形式嵌入到分割網絡中，以增加提取特徵的動態表示能力。

In the EIM module, the expertise level cues of multiple raters are formed as a normalized expertness vector V

2 R1×1×N, where N represents the total number of raters and ΣN i=1 Vi = 1.
在 EIM 模塊中，多個評估者的專業水平線索形成為歸一化的專業向量 V 2 R1×1×N，其中 N 表示評估者的總數，∑N i=1 Vi = 1。

It is fed to the network as prior knowledge and determines the actual soft GT labels that are set as the network’s target.
它作為先驗知識被饋送到網絡，並確定設置為網絡目標的實際軟 GT 標籤。

Specifically, the soft GT label used in the training is determined by the annotations of individual raters multiplied by their corresponding weight in the expertness vector V , which is denoted as:
具體來說，訓練中使用的軟 GT 標籤由各個評分者的註釋乘以他們在專家向量 V 中的相應權重確定，表示為：

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

where φ denotes the model parameters; x is the input image; and Si means the annotation mask by the ith expert.
其中 φ 表示模型參數； x 是輸入圖像； Si 表示第 i 個專家的註釋掩碼。

During each training iteration, the expertness vector V is dynamically set with three different strategies alternatively, including the majority vote mode (i.e., uniform weight among all raters), single rater mode (i.e., assign weight of 1 to single random rater and suppress the rest raters to 0), and random weight assignment (i.e., assign each rater’s weight randomly and then normalize to a unit vector).
在每次訓練迭代中，專家向量 V 交替使用三種不同的策略動態設置，包括多數投票模式（即所有評分者的權重一致）、單一評分者模式（即為單個隨機評分者分配權重為 1 並抑制 將評分者重置為 0）和隨機權重分配（即隨機分配每個評分者的權重，然後歸一化為單位向量）。

By using different strategies to assign the expertness vector, the model learns to associate the influence/weight of individual raters on the final soft predictions.
通過使用不同的策略來分配專家向量，該模型學習將各個評估者的影響/權重對最終軟預測進行關聯。

In addition, the dynamic ex-pertness vector together with the adaptively changing GT label works as an effective data augmentation strategy that increases the data variability and input-output data pairs being fed to the model.
此外，動態專家向量與自適應變化的 GT 標籤一起作為一種有效的數據增強策略，增加了數據可變性和輸入到模型的輸入輸出數據對。

In the inference stage, only the majority vote mode is used by default to set the expertness vector, making it easily applicable for clinical applications.

在推理階段，默認僅使用多數投票模式來設置專家向量，使其易於應用於臨床應用。

In order to integrate the multi-rater expertise cues into the semantic feature representation effectively, we utilize a ConvLSTM module [40] to generate the enhanced features embedded with the expertness vector as hidden state, as shown in Fig. 2(b).
為了將多評價者的專業知識線索有效地整合到語義特徵表示中，我們利用 ConvLSTM 模塊 [40] 生成嵌入專家向量作為隱藏狀態的增強特徵，如圖 2（b）所示。

ConvLSTM is a powerful recurrent model that not only captures the correlation between features and different expertise levels (i.e., the hidden state), but also summarizes the discriminative dynamic features.
ConvLSTM 是一個強大的循環模型，它不僅可以捕獲特徵與不同專業知識水平（即隱藏狀態）之間的相關性，還可以總結具有判別力的動態特徵。

To be more specific, we take the feature map from the bottleneck layer (i.e., f5) as input to the proposed EIM and use the normalized expertness vector V ∈ R1×1×N as initial hidden state h0.
更具體地說，我們將瓶頸層（即 f5）的特徵圖作為所提出的 EIM 的輸入，並使用歸一化的專家向量 V ∈ R1×1×N 作為初始隱藏狀態 h0。

To transfer the expertness vector into a proper format for ConvLSTM, we expand V to the same dimension as that of f5.
為了將專家向量轉換為適合 ConvLSTM 的格式，我們將 V 擴展到與 f5 相同的維度。

The procedure can be defined as:
該過程可以定義為：

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

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

其中 t 表示 ConvLSTM 中的時間步長，t 表示時間 t 的迭代過程。

After T steps, which is empirically set as two in this work, an enhanced feature f5e = hT embedded with expertise cues is generated.
在本工作中根據經驗設置為 2 的 T 步之後，生成了嵌入專業知識線索的增強特徵 f5e = hT。

The enhanced f5e is further sent to the U-Net decoder to obtain the coarse calibrated prediction P1 and decoded feature F1.
增強後的 f5e 進一步發送到 U-Net 解碼器，得到粗校準預測 P1 和解碼特徵 F1。

3.4. Multirater 多評級者

Reconstruction Module 重建模塊

In order to further enhance the association between the expertness vector with the model prediction, and to capture the valuable inter-rater disagreement cues, a Multi-rater Reconstruction Module (MRM) is proposed to reconstruct the individual rater’s annotation from the corresponding soft prediction P1 and the given expertness vector V .
為了進一步增強專家向量與模型預測之間的關聯，並捕獲有價值的評分者間分歧線索，提出了多評分者重建模塊（MRM）從相應的軟預測 P1 重建個體評分者的註釋 和給定的專業度向量 V 。

Based on the reconstructed multi-rater’s annotation, an uncertainty map that reflects the inter-rater variability is generated.
基於重建的多評價者註釋，生成反映評價者間變異性的不確定性圖。

Specifically, as shown in Fig. 2(c), the initial prediction P1 and the input image are concatenated and fed into an encoder-decoder network with VGG16 [41] as the feature encoder, since VGG architecture is well known for its superior capability that preserves the topological and perceptual features of the input image [22, 33].
具體來說，如圖 2(c) 所示，初始預測 P1 和輸入圖像被連接並饋入以 VGG16 [41] 作為特徵編碼器的編碼器-解碼器網絡，因為 VGG 架構以其卓越的能力而聞名 它保留了輸入圖像的拓撲和感知特徵 [22, 33]。

The corresponding ex-pertness vector V is applied at the bottleneck layer of the MRM via another EIM module.
相應的專業向量 V 通過另一個 EIM 模塊應用於 MRM 的瓶頸層。

The decoder of MRM tries to reconstruct the annotations of individual raters via multiple 1×1 convolution layers (i.e.,Conv1×1) in the last layer.
MRM 的解碼器嘗試通過最後一層中的多個 1x1 卷積層（即 Conv1x1）重建單個評分者的註釋。

Here, we employ a reconstruction loss, lossrec, to measure the extent to which the reconstructed multi-raters’ grading is similar to that of the real annotation marked by individual raters, which is defined as lossrec =1/N ΣN i=1 LBCE(Si, S¯i).
在這裡，我們使用重建損失 lossrec 來衡量重建後的多評估者的評分與單個評估者標記的真實註釋的相似程度，定義為 lossrec =1/N ∑N i=1 LBCE (Si, S¯i)。 .

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

Here LBCE denotes the binary cross entropy loss; N is the total number of experts; Si and Si¯ ∈ RW×H×C denote the annotation marked by the ith expert and the corresponding reconstructed prediction; W, H, and C denote the image width, height, and the number of channels, respectively.

這裡 LBCE 表示二元交叉熵損失； N 是專家總數； Si 和 Si¯ ∈ RW×H×C 表示第 i 個專家標註的標註和對應的重構預測； W、H 和 C 分別表示圖像寬度、高度和通道數。

To further improve the reconstruction performance of the MRM module, the fused soft label GTsoft = ΣN i (Vi, Si), together with the given expertness vector, is also fed into the network to reconstruct the individual rater's grading.
為了進一步提高 MRM 模塊的重建性能，融合的軟標籤 GTsoft = ∑N i (Vi, Si) 與給定的專家向量一起也被輸入網絡以重建個體評分者的評分。

A consistency loss, losscon, is proposed to enhance the coherence between the features extracted from the soft prediction P1 and GTsoft, as losscon = 1/K ΣK i=1 (1/2)(||Di − D¯i ||)^2.
提出了一致性損失 losscon 來增強從軟預測 P1 和 GTsoft 中提取的特徵之間的一致性，因為 losscon = 1/K ΣK i=1 (1/2)(||Di − D¯i ||) ^2。

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

Here {D¯i}Ki =1 and {Di}Ki =1 represent feature sets extracted from the encoder by using P1 and GTsoft as input, respectively; K indicates the number of convolutional blocks where the features are extracted from and for the VGG16 backbone K = 5.
這裡{D¯i}Ki =1 和{Di}Ki =1 表示分別使用 P1 和 GTsoft 作為輸入從編碼器中提取的特徵集； K 表示從 VGG16 主幹 K = 5 中提取特徵的捲積塊的數量。

After reconstructing the individual rater's grading via the MRM, the uncertainty map of grading inconsistency can be estimated via the pixel-wise standard deviation of the multiple rater's predictions, using:
在通過 MRM 重建單個評分者的評分後，評分不一致的不確定性圖可以通過多個評分者預測的像素標準偏差來估計，使用：

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

The obtained uncertainty map is sent to the next module to further refine the initial coarse prediction P1.
將獲得的不確定性圖發送到下一個模塊以進一步細化初始粗略預測 P1。

3.5. Multirater 多評級者

Perception Module 感知模塊

The grading inconsistency among multiple experts, i.e., the inter-rater variability, reflects the uncertainty or

difficulty levels of different regions across the medical image.
多個專家之間的分級不一致，即評分者間的可變性，反映了醫學圖像不同區域的不確定性或難度級別。

Thus, how to better take advantage of this information to further improve the segmentation performance is an important research problem.
因此，如何更好地利用這些資訊進一步提高分割性能是一個重要的研究問題。

In this paper, we innovatively design a Multi-rater Perception Module (MPM), which can better capture and emphasize ambiguous regions by using the designed multi-branch soft attention mechanism.

在本文中，我們創新地設計了一個多評估者感知模塊（MPM），它可以通過使用設計的多分支軟注意力機制更好地捕捉和強調模糊區域。

Given the feature map F1 obtained by the U-Net backbone and the estimated uncertainty map Umap obtained by the MRM, we use a spatial attention strategy [51] to emphasize the highly uncertain regions.
鑑於 U-Net 主幹獲得的特徵圖 F1 和 MRM 獲得的估計不確定性圖 Umap，我們使用空間注意策略 [51] 來強調高度不確定的區域。

However, the estimated uncertainty map might contain potential inaccuracy or incompleteness near the object boundaries, which may negatively affect the model performance if a 'hard' spatial attention is used.
然而，估計的不確定性圖可能包含對象邊界附近的潛在不准確或不完整，如果使用"硬"空間注意力，這可能會對模型性能產生負面影響。

Therefore, we employ a 'soft' attention which aims to enlarge the coverage area of the uncertain regions, so as to effectively perceive and capture the disagreement cues among multiple raters.
因此，我們採用了一種"軟"注意力，旨在擴大不確定區域的覆蓋範圍，從而有效地感知和捕捉多個評分者之間的分歧線索。

The soften operation can be formulated by:
軟化操作可以表示為：

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

where FGauss indicates a convolution operation with a Gaussian kernel k and zero bias, and Ωmax indicates a maximum function to preserve the higher values between the Gaussian filtered map and the original uncertainty map Umap.

其中 FGauss 表示具有高斯核 k 和零偏差的捲積運算，Ωmax 表示最大函數以保留高斯濾波圖和原始不確定性圖 Umap 之間的較高值。

In this paper, the size and standard deviation of the Gaussian kernel k are learnable through the model training procedure and initialized with 32 and 4, respectively.
在本文中，高斯核 k 的大小和標準偏差可通過模型訓練過程學習，並分別用 32 和 4 初始化。

Apart from the highly uncertain regions, the soft attention mechanism is applied on the initial coarse prediction map P1 as well to enhance the highly certain regions for feature map F1.
除了高度不確定的區域外，還對初始粗略預測圖 P1 應用了軟注意力機制，以增強特徵圖 F1 的高度確定性區域。

In other words, both highly uncertain and certain regions are strengthened for F1.
換句話說，F1 的高度不確定性和某些區域都得到了加強。

For joint optic cup and disc segmentation task, F1 is sent to four parallel branches with soft spatial attentions obtained from Aj = {Ucupmap, Udiscmap, P1cup, P1disc}4j=1, as shown in Fig. 2(c).
對於聯合視杯和視盤分割任務，F1 被發送到四個並行分支，從 Aj = {Ucupmap, Udiscmap, P1cup, P1disc}4j=1 獲得軟空間注意力，如圖 2(c) 所示。

A skip connection is adopted between the original feature F1 and the spatially enhanced features, so as to alleviate potential errors in the attention map being propagated to the network.
在原始特徵 F1 和空間增強特徵之間採用跳躍連接，以減輕傳播到網絡的注意力圖的潛在錯誤。

The procedure is described as:
程序描述如下：

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

where ⊗ denotes the pixel-wise multiplication operation and F˜j represents the refined feature from the jth branch using the soft attention operation.
其中 ⊗ 表示逐像素乘法運算，F~j 表示使用軟注意力操作從第 j 個分支中精煉的特徵。

The refined feature sets are further concatenated and fed to a Conv1×1 layer to obtain the final segmentation prediction M, as in:
細化的特徵集進一步連接並饋送到 Conv1x1 層以獲得最終的分割預測 M，如下所示：

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

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

最後，所提出的 MRNet 框架的總訓練損失 L 是 U-Net 主幹、MRM 模塊和 MPM 模塊的損失的組合，可以表示為：

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

where LBCE denotes the binary cross entropy loss;
其中 LBCE 表示二元交叉熵損失；

α is a hyper-parameter that balances the weight of reconstruction loss lossrec and consistency loss losscon in the MRM module and empirically set as 0.7 in this work.
α 是一個超參數，用於平衡 MRM 模塊中重建損失 lossrec 和一致性損失 losscon 的權重，並在本工作中根據經驗設置為 0.7。

4. Experiments 實驗

4.1. Datasets 數據集

Extensive experiments are conducted to verify the effectiveness of the proposed framework on five different types of medical segmentation tasks with data from varied image modalities, including color fundus images, CT and MRI.
進行了廣泛的實驗，以驗證所提出的框架在五種不同類型的醫學分割任務上的有效性，這些任務來自不同圖像模式的數據，包括彩色眼底圖像、CT 和 MRI。

RIGA benchmark [1] is a publicly available dataset for retinal cup and disc segmentation, which contains in total of 750 color fundus images from three sources, including 460 images from MESSIDOR, 195 images from BinRushed and 95 images from Magrabia.
RIGA 基準 [1] 是用於視網膜杯盤分割的公開數據集，它包含來自三個來源的總共 750 張彩色眼底圖像，包括來自 MESSIDOR 的 460 張圖像、來自 BinRushed 的 195 張圖像和來自 Magrabia 的 95 張圖像。

Six glaucoma experts from different organizations labeled the optic cup and disc contour masks manually for the RIGA benchmark [1].
來自不同組織的六名青光眼專家為 RIGA 基準手動標記視杯和視盤輪廓掩模 [1]。

During model training, we select 195 samples from BinRushed and 460 samples from MESSIDOR as the training set, following[53].
在模型訓練過程中，我們從 BinRushed 中選擇 195 個樣本和 MESSIDOR 中的 460 個樣本作為訓練集，如下[53]。

The Magrabia set with 95 samples is selected as the test set to evaluate the model, which is not homologous to the training dataset.
選擇具有 95 個樣本的 Magrabia 集作為測試集來評估模型，該集與訓練數據集不同源。

Parameters of the U-Net encoder are initialized with the model pre-trained on ImageNet [27].
U-Net 編碼器的參數使用在 ImageNet [27] 上預訓練的模型進行初始化。

QUBIQ benchmark [32], namely Quantification of Uncertainties in Biomedical Image Quantification Challenge, is a recently available challenge dataset specifically for the evaluation of inter-rater variability.
QUBIQ 基準 [32]，即生物醫學圖像量化挑戰中的不確定性量化，是最近可用的挑戰數據集，專門用於評估評估者間的可變性。

QUBIQ contains four different segmentation datasets with CT and MRI modalities, including brain growth (one task, MRI, seven raters, 34 cases for training and 5 cases for testing), brain tumor (one task, MRI, three raters, 28 cases for training and 4 cases for testing), prostate (two subtasks, MRI, six raters, 33 cases for training and 15 cases for testing), kidney (one task, CT, three raters, 20 cases for training and 4 cases for testing).
QUBIQ 包含四個不同的具有 CT 和 MRI 模式的分割數據集，包括大腦生長（一項任務，MRI，7 個評估者，34 個訓練案例和 5 個測試案例），腦腫瘤（一項任務，MRI，三個評估者，28 個訓練案例）和測試 4 例），前列腺（兩個子任務，MRI，六個評分者，訓練 33 例，測試 15 例），腎臟（一個任務，CT，三個評分者，訓練 20 例，測試 4 例）。

4.2. Experimental Setup 實驗裝置

4.2.1 Implementation Details 實施細則

In our experiments, the main framework utilizes the U-Net architecture with ResNet34 as the backbone, and the MRM module utilizes the DeepLab-V3+ architecture with VGG- 16 as the backbone.
在我們的實驗中，主要框架採用以 ResNet34 為骨幹的 U-Net 架構，MRM 模塊採用以 VGG-16 為骨幹的 DeepLab-V3+ 架構。

The network is implemented with the PyTorch platform and trained/tested on a Tesla P40 GPU with 24GB of memory.
該網絡使用 PyTorch 平台實現，並在具有 24GB 內存的 Tesla P40 GPU 上進行訓練/測試。

All training and test images are uniformly resized to the dimension of 256×256 pixels.
所有訓練和測試圖像都統一調整為 256×256 像素的尺寸。

The proposed network is trained in an end-to-end manner using the Adam optimizer [25], and it takes about 4 hours to train our model with a mini-batch size of 8 for 60 epochs.
提出的網絡使用 Adam 優化器 [25] 以端到端的方式進行訓練，用 8 的 mini-batch 大小訓練我們的模型大約需要 4 個小時，持續 60 個 epoch。

The initial learning rate is set to 1 × 10^(−4).
初始學習率設置為 1 × 10^(−4)。

4.2.2 Evaluation Metric 評估指標

The target of the proposed network is to produce probability mapMthat can reflect the underlying inter-rater agreement/ disagreement, i.e., calibrated predictions, for medical image segmentation. In order to better evaluate the calibrated model predictions, we use soft dice coefficient(D) / Intersection Over Union (IoU) metrics through multiple threshold levels, set as (0.1, 0.3, 0.5, 0.7, 0.9) in this paper, instead of using a single threshold (e.g., 0.5).
所提出的網絡的目標是產生概率圖 M，它可以反映潛在的評估者間的一致/不一致，即校準預測，用於醫學圖像分割。 為了更好地評估校准後的模型預測，我們通過多個閾值級別使用軟骰子係數（D）/交集交叉（IoU）指標，在本文中設置為 (0.1, 0.3, 0.5, 0.7, 0.9)，而不是 使用單個閾值（例如，0.5）。

At each threshold level, the predicted probability map M and soft GT GTsoft are binarized with the given threshold and then the D and IoU metrics are computed.
在每個閾值級別，預測概率圖 M 和軟 GT GTsoft 用給定的閾值進行二值化，然後計算 D 和 IoU 度量。

The D and IoU scores obtained at multiple thresholds are averaged and then we obtain the soft metrics, denoted as Ds and IoUs, respectively.
在多個閾值處獲得的 D 和 IoU 分數被平均，然後我們獲得軟度量，分別表示為 Ds 和 IoU。

The higher the soft scores, the better calibrated the model performance.
軟分數越高，模型性能校準得越好。

4.3. Experimental Results 實驗結果

4.3.1 Performance of the Multi-rater Strategy
多評估者策略的表現

In order to verify that our multi-rater strategy can generate better calibrated segmentation maps under different given expertness conditions, we conduct quantitative experiments with different expertness setups on the RIGA test set in Table 2.
為了驗證我們的多評估者策略可以在不同給定的專業條件下生成更好的校準分割圖，我們在表 2 中的 RIGA 測試集上使用不同的專業設置進行了定量實驗。

Here, M1-M6 refer to the U-Net baseline model trained with the corresponding labels graded by Raters 1-6, respectively.
在這裡，M1-M6 是指分別用評分者 1-6 分級的相應標籤訓練的 U-Net 基線模型。

In addition, three commonly used multirater strategies are employed to train the U-Net baseline model, including majority vote (i.e., U-Net baseline model [38] trained with the GT labels obtained by majority vote), label sampling [19] and multi-head strategies [16], denoted as MV-UNet, LS-UNet and MH-UNet, respectively, in Table 2.
此外，採用三種常用的多評級策略來訓練 U-Net 基線模型，包括多數投票（即使用多數投票獲得的 GT 標籤訓練的 U-Net 基線模型 [38]）、標籤採樣 [19] 和 多頭策略 [16]，分別表示為 MV-UNet、LS-UNet 和 MH-UNet，見表 2。

The performance of the comparison models is evaluated against various GT labels generated from different expertness vectors, including single rater condition (Raters1-6 raw gradings), random condition, average weight condition and STAPLE [50].
比較模型的性能是針對從不同專家向量生成的各種 GT 標籤進行評估的，包括單一評估者條件（Raters1-6 原始分級）、隨機條件、平均重量條件和 STAPLE [50]。

For the random condition, we randomly select three groups of results under different random expert-ness and report the average performance.
對於隨機條件，我們在不同的隨機專家度下隨機選擇三組結果並報告平均性能。

As listed in Table 2, the proposed MRNet consistently achieves superior performance under different conditions, reflecting the dynamic representation capability of the MRNet by incorporating the expertise cues of individual raters.
如表 2 中所列，所提出的 MRNet 在不同條件下始終如一地實現了卓越的性能，反映了 MRNet 通過結合各個評估者的專業知識線索的動態表示能力。

In addition, it is worth noting that our approach achieves the best performance under majority vote (i.e., average weight condition), with a large performance margin over all the other models, including the MV-UNet which is specifically trained with the majority vote consensus labels.
此外，值得注意的是，我們的方法在多數投票（即平均權重條件）下實現了最佳性能，與所有其他模型相比具有較大的性能差距，包括專門使用多數投票共識訓練的 MV-UNet 標籤。

These empirical experiments demonstrate the effectiveness of the proposed framework which is tailored for medical image segmentations with multi-rater annotations, by taking advantage of the proposed dynamic expertness inferring and multi-rater agreement modeling.
這些實證實驗證明了所提出的框架的有效性，該框架通過利用所提出的動態專家推斷和多評分者協議建模為具有多評分者註釋的醫學圖像分割量身定制。

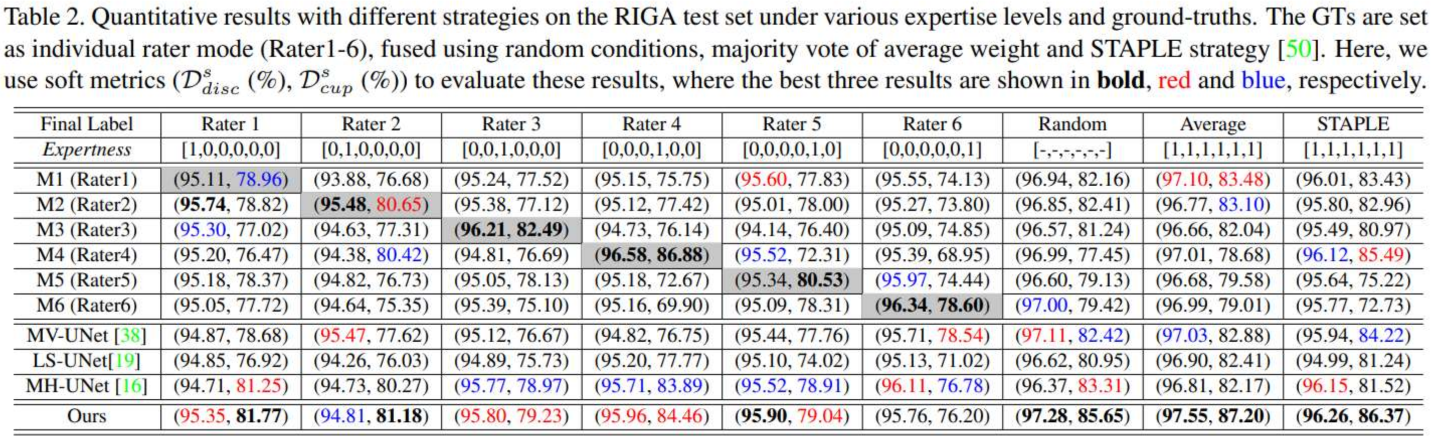


Table 2. Quantitative results with different strategies on the RIGA test set under various expertise levels and ground-truths. The GTs are set as individual rater mode (Rater1-6), fused using random conditions, majority vote of average weight and STAPLE strategy [50]. Here, we use soft metrics ($\mathcal{D}^s_{disc}$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

Table 2. Quantitative results with different strategies on the RIGA test set under various expertise levels and ground-truths.
表 2. RIGA 測試集在不同專業水平和真實情況下不同策略的定量結果。

The GTs are set as individual rater mode (Rater1-6), fused using random conditions, majority vote of average weight and STAPLE strategy [50].
GT 被設置為單獨的評估者模式（Rater1-6），使用隨機條件、平均權重的多數投票和 STAPLE 策略進行融合 [50]。

Here, we use soft metrics (Dsdisc (%), Dscup (%)) to evaluate these results, where the best three results are shown in bold, red and blue, respectively.
在這裡，我們使用軟指標（Dsdisc (%)、Dscup (%)）來評估這些結果，其中最好的三個結果分別以粗體、紅色和藍色顯示。

4.3.2 Comparisons with State-of-the-arts 與最先進技術的比較

To demonstrate the advantage of the proposed MRNet, we compare our method with the state-of-the-art (SOTA) methods for joint optic cup and disc segmentation task.
為了證明所提出的 MRNet 的優勢，我們將我們的方法與用於聯合視杯和椎間盤分割任務的最先進（SOTA）方法進行了比較。

We use the publicly released code with default parameters to retrain the SOTA methods, with the same training/test set as that of ours for a fair comparison.
我們使用公開發布的帶有默認參數的代碼來重新訓練 SOTA 方法，使用與我們相同的訓練/測試集進行公平比較。

Table 3 quantitatively compares our framework with five SOTA cup/disc segmentation methods, including ResUnet[53], CENet [15], AGNet [58], BEAL [45] and pOSAL [46] on the RIGA test set.
表 3 在 RIGA 測試集上定量比較了我們的框架與五種 SOTA 杯/盤分割方法，包括 ResUnet[53]、CENet[15]、AGNet[58]、BEAL[45]和 pOSAL[46]。

As shown in Table 3, our proposed MRNet consistently achieves superior performance compared with SOTA optic cup/disc segmentation methods.
如表 3 所示，與 SOTA 視杯/視盤分割方法相比，我們提出的 MRNet 始終具有卓越的性能。

The performance improvement is especially prominent for the retinal cup segmentation where the inter-observer variability is more significant, with an increase of 1.2% for soft dice coefficient value over the current best method.
對於觀察者間變異性更顯著的視網膜杯分割，性能改進尤為突出，軟骰子係數值比當前最佳方法增加了 1.2%。

Fig. 4 shows two typical examples generated by our MRNet and other SOTA methods. It is obvious that the probability map generated by the proposed model is better calibrated compared with other methods, especially for the ambiguous regions among different experts.
圖 4 顯示了我們的 MRNet 和其他 SOTA 方法生成的兩個典型示例。 很明顯，與其他方法相比，所提出的模型生成的概率圖具有更好的校準能力，尤其是對於不同專家之間的模糊區域。

Thus, the predictions generated by the proposed MRNet is able to better reflect the underlying dis-/agreement among multiple experts.
因此，所提出的 MRNet 生成的預測能夠更好地反映多個專家之間的潛在分歧/分歧。

4.3.3 Ablation Studies 消融研究

In this section, ablation studies are performed over each component of the proposed MRNet, including the EIM, MRM and MPM, as listed in Table 4 and Table 5.
在本節中，對提議的 MRNet 的每個組件進行消融研究，包括 EIM、MRM 和 MPM，如表 4 和表 5 中所列。

All experiments are evaluated using the soft GT obtained with majority vote, i.e., the average weight expertness condition.
所有實驗都使用通過多數投票獲得的軟 GT 進行評估，即平均權重專家條件。

In Table 4, as we sequentially adding the proposed modules on top of the U-Net baseline, the model performance is gradually improved, especially for that of the optic cup.
在表 4 中，隨著我們在 U-Net 基線之上依次添加所提出的模塊，模型性能逐漸提高，尤其是對於視杯的性能。

Firstly, by integrating the EIM with ConvLSTM into the UNet baseline, the Dsc up value is increased by 1.0%.
首先，通過將帶有 ConvLSTM 的 EIM 集成到 UNet 基線中，Dsc up 值增加了 1.0%。

Compared to the direct condition operation by concatenating ex-pertness with feature maps, the EIM with ConvLSTM operation achieves better performance (Table 4(b) vs. (c)).

與通過將專家與特徵圖連接的直接條件操作相比，具有 ConvLSTM 操作的 EIM 實現了更好的性能（表 4（b）與（c））。

This indicates that the introduction of multi-rater expertise knowledge via EIM with ConvLSTM improves the dynamic representation capability of the model and the exploitation of multi-rater annotations can arrive at better calibrated predictions.
這表明通過帶有 ConvLSTM 的 EIM 引入多評價者專業知識提高了模型的動態表示能力，並且多評價者註釋的利用可以達到更好的校準預測。

Additionally, in order to effectively utilize the multi-rater cues for calibrating the segmentation results, the MRM and MPM modules are specifically designed to reconstruct the raw multi-rater gradings and further to utilize the multi-rater (dis-)agreement cues, which boosts the Dscup metric by 2.0% and 1.5%, respectively.
此外，為了有效地利用多評價者線索來校準分割結果，MRM 和 MPM 模塊專門設計用於重建原始多評價者評分，並進一步利用多評價者（不）同意線索， 分別將 Dscup 指標提高了 2.0% 和 1.5%。

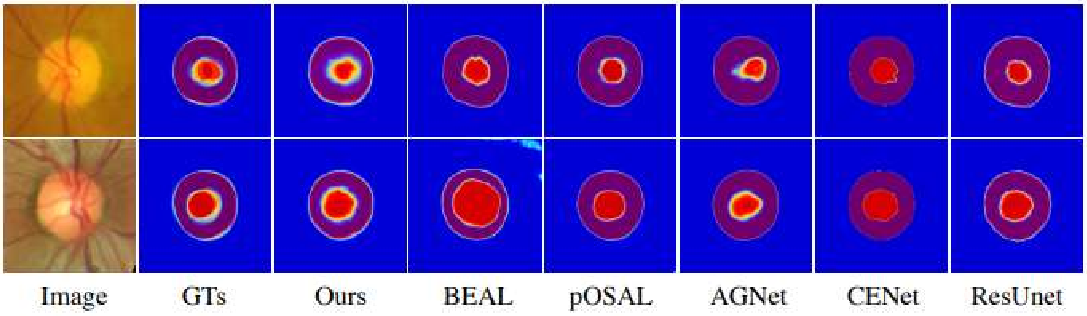


Figure 4. Visual comparisons of our MRNet with the state-of-the-arts for joint optic cup and disc segmentation tasks.

Figure 4. Visual comparisons of our MRNet with the state-of-thearts for joint optic cup and disc segmentation tasks.
圖 4. 我們的 MRNet 與用於聯合視杯和視盤分割任務的最新技術的視覺比較。

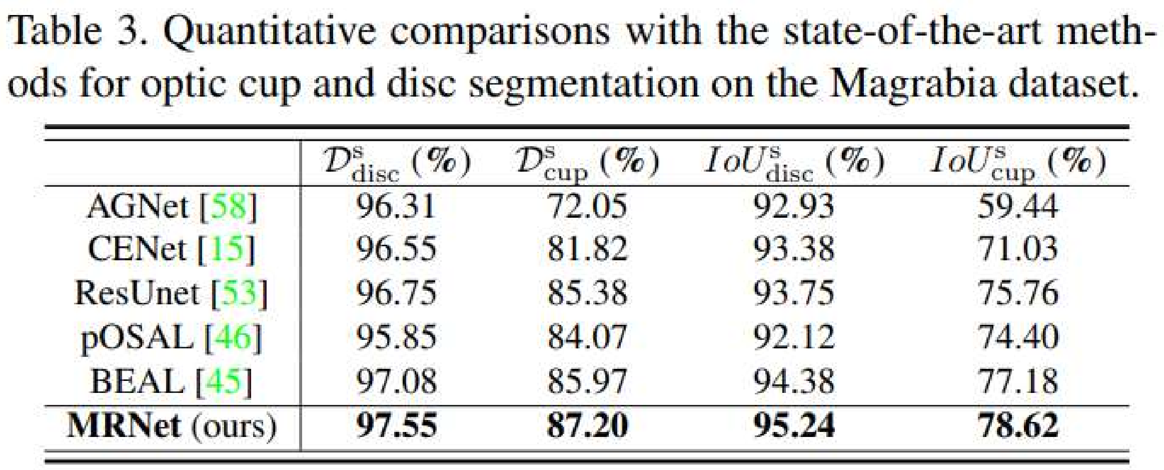


Table 3. Quantitative comparisons with the state-of-the-art methods for optic cup and disc segmentation on the Magrabia dataset.

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

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

the Magrabia dataset.
表 3. 在 Magrabia 數據集上與最先進的視杯和視盤分割方法的定量比較。

To further investigate the influence of individual losses and operations in the MRM and MPM, i.e., the MAM module, a set of ablation studies is conducted for the reconstruction loss (lossrec), consistency loss (losscon) and soft atten-tion operation (Softop).
為了進一步研究 MRM 和 MPM 中單個損失和操作的影響，即 MAM 模塊，對重建損失 (lossrec)、一致性損失 (losscon) 和軟注意力操作 (Softop) 進行了一組消融研究)。

In Table 5, the reconstruction loss significantly improves the Dsc up by 1.2%, reflecting the necessity of reconstructing raw multi-rater's grading from the fused soft GT/prediction.
在表 5 中，重建損失使 Dsc 顯著提高了 1.2%，反映了從融合的軟 GT/預測重建原始多評級者分級的必要性。

By adding the consistency loss to constrain the features extracted from soft GT and coarse prediction, the Dsc up is further improved by 0.8%.
通過添加一致性損失來約束從軟 GT 和粗略預測中提取的特徵，Dsc up 進一步提高了 0.8%。

Moreover, comparing Table 5 (iv) and (v), the soft attention operation further boosts the Dsc up by 1.2% compared with using 'hard' attention, achieving the final Dsc up score of 87.2%.
此外，比較表 5 (iv) 和 (v)，與使用"硬"注意力相比，軟注意力操作進一步將 Dsc 提升了 1.2%，最終達到了 87.2% 的 Dsc up 分數。

This verifies that the proposed soft attention mechanism can better emphasize both certain and uncertain regions and further to improve the calibration performance of the model.
這驗證了所提出的軟注意力機制可以更好地強調某些區域和不確定區域，並進一步提高模型的校準性能。

We also investigate the influence of different U-Net backbones and the hyper-parameter.
我們還研究了不同 U-Net 主乾和超參數的影響。

With a stronger backbone (ResNet101 vs. ResNet34), the model performance can be further improved (88.45% vs 87.20%), in terms of Dscup(%).
憑藉強大的主幹（ResNet 101 vs. ResNet 34），模型性能可以進一步提高（88.45% vs 87.20%），就 D'Cup(%) 而言。

When is set as 0.3, 0.7, 0.9, the corresponding Dscup(%) is 86.17%, 87.20% and 86.87%, respectively.
當設置為 0.3、0.7、0.9 時，對應的 Dscup(%)分別為 86.17%、87.20%和 86.87%。

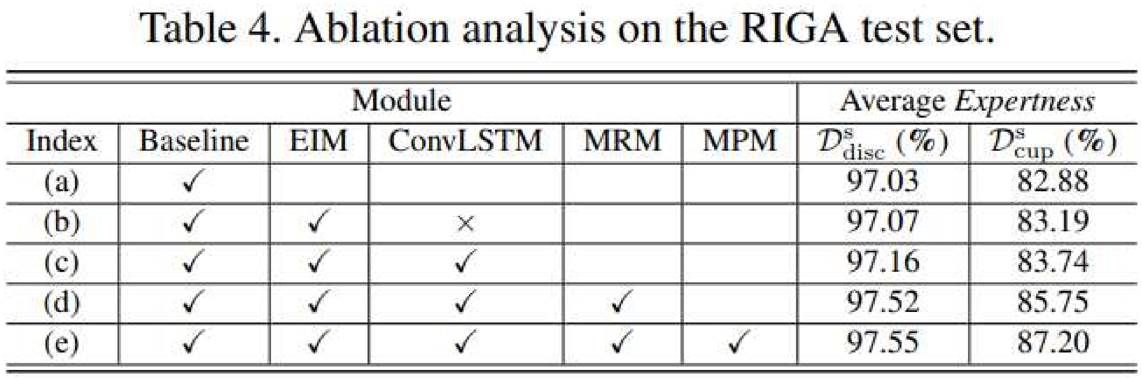


Table 4. Ablation analysis on the RIGA test set.

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

Table 4. Ablation analysis on the RIGA test set.
表 4. RIGA 測試集的消融分析。

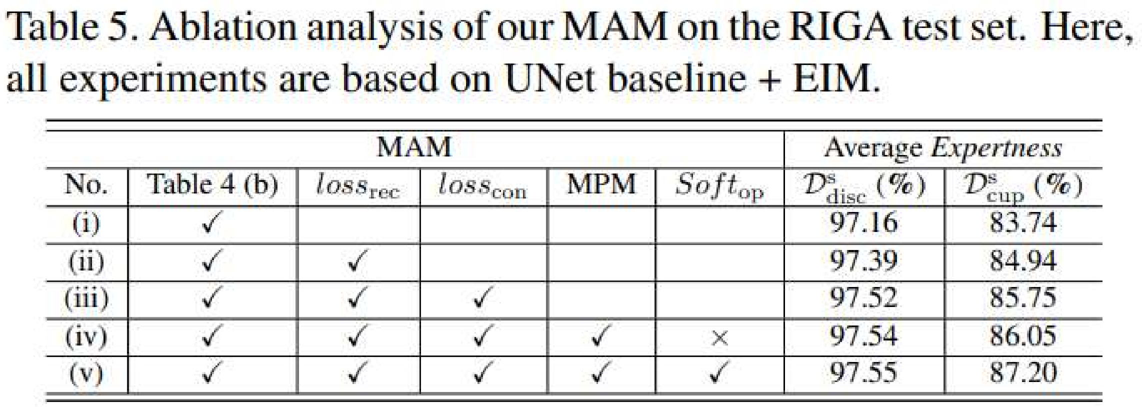


Table 5. Ablation analysis of our MAM on the RIGA test set. Here, all experiments are based on UNet baseline + EIM.

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

Table 5. Ablation analysis of ourMAMon the RIGA test set. Here, all experiments are based on UNet baseline + EIM.
表 5. 我們的 MAMon RIGA 測試集的消融分析。 在這裡，所有的實驗都是基於 UNet 基線 + EIM。

4.3.4 Generalization Capability 泛化能力

To further verify the effectiveness and generalization capability of the proposed MRNet, a generalization experiment is conducted on a recently released QUBIQ dataset for four types of medical image segmentation tasks that contain both CT and MRI modalities.
為了進一步驗證所提出的 MRNet 的有效性和泛化能力，在最近發布的 QUBIQ 數據集上進行了泛化實驗，用於包含 CT 和 MRI 模態的四種醫學圖像分割任務。

Several commonly used multi-rater strategies are adopted for comparison, including U-Net [38] based on majority vote (MV-UNet), label sampling [19] (LS-UNet) and multiple head strategies [16] (MH-UNet).
採用幾種常用的多評級策略進行比較，包括基於多數投票的 U-Net [38] (MV-UNet)、標籤採樣 [19] (LS-UNet) 和多頭策略 [16] (MH-UNet) ）。

As listed in Table 6, the proposed MRNet achieves better calibrated performance compared with other commonly used methods and multi-rater strategies.
如表 6 中所列，與其他常用方法和多評級策略相比，所提出的 MRNet 實現了更好的校準性能。

These quantitative results again verify that the underlying agreement/ disagreement information among multiple experts regarding the pathological region are beneficial to improve calibrated segmentation accuracy through our multi-rater agreement modeling.
這些定量結果再次驗證了多個專家之間關於病理區域的基本一致/不一致信息有利於通過我們的多評估者一致建模提高校準分割的準確性。

Several representative examples of the comparison methods for four different types of medical image segmentation are visualized in Fig. 5.
四種不同類型醫學圖像分割的比較方法的幾個代表性示例如圖 5 所示。

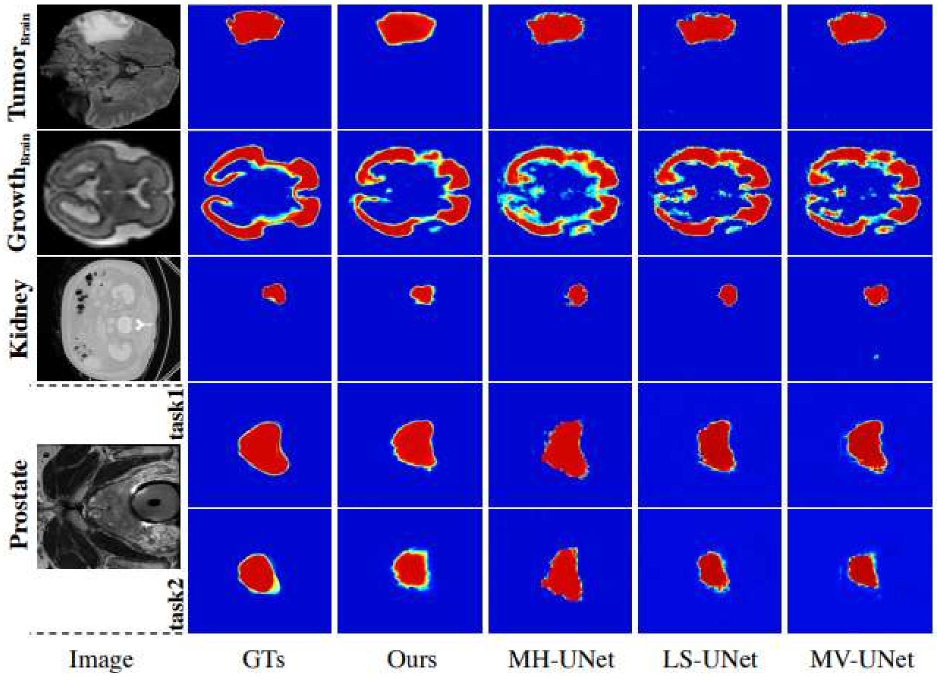


Figure 5. Segmentation results of different strategies for four different medical segmentation tasks on the QUBIQ dataset.

Figure 5. Segmentation results of different strategies for four different medical segmentation tasks on the QUBIQ dataset.
圖 5. QUBIQ 數據集上四種不同醫學分割任務的不同策略的分割結果。

Table 6. Quantitative evaluation of five medical segmentation subtasks with multi-rater modeling on the QUBIQ dataset, including the segmentation of kidney (Dskidney), brain growth (Dsbrain), brain tumor (Dstumor) and two prostate tasks (Dspros1 and Dspros2).
表 6. 在 QUBIQ 數據集上使用多評估者建模對五個醫學分割子任務的定量評估，包括腎臟 (Dskidney)、腦生長 (Dsbrain)、腦腫瘤 (Dstumor) 和兩個前列腺任務 (Dspros1 和 Dspros2) 的分割。

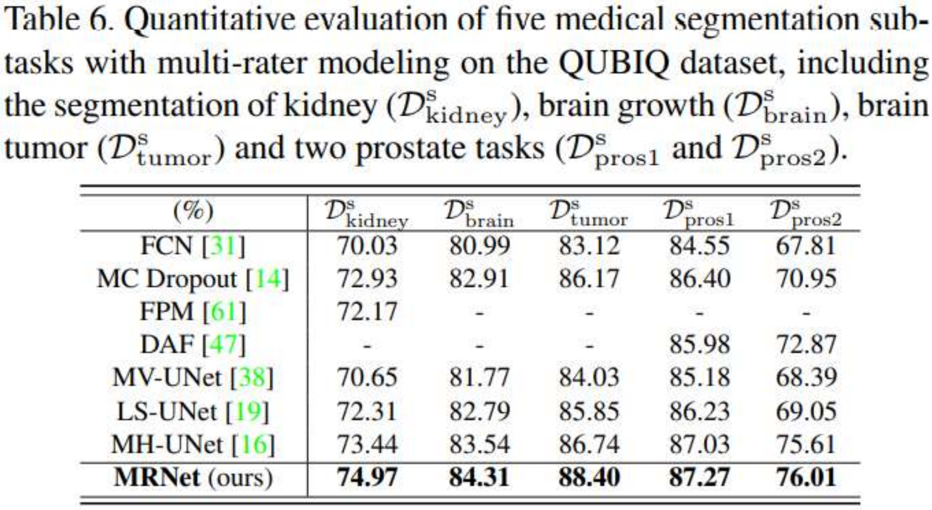


Table 6. Quantitative evaluation of five medical segmentation subtasks with multi-rater modeling on the QUBIQ dataset, including the segmentation of kidney ($\mathcal{D}^{\text{s}}_{\text{kidney}}$), brain growth ($\mathcal{D}^{\text{s}}_{\text{brain}}$), brain tumor ($\mathcal{D}^{\text{s}}_{\text{tumor}}$)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

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

5. Conclusion 結論

In this work, we focus on the utilization of rich annotation information from multiple experts, which are relatively less-explored but widely presented in the medical image grading procedure.
在這項工作中，我們專注於利用來自多個專家的豐富註釋資訊，這些資訊相對較少探索但在醫學圖像分級過程中廣泛存在。

We proposed to incorporate the multirater (dis-)agreement cues in our MRNet framework and generate calibrated model predictions that better reflected the underlying agreement among multiple experts.
我們提議在我們的 MRNet 框架中加入多評價者（不一致）同意線索，並生成校準模型預測，以更好地反映多位專家之間的基本一致。

This was achieved by the use of an expertise-aware inferring module to explicitly integrate graders expertise cues into high-level semantic features, as well as a multi-rater agreement modeling module to reconstruct gradings of individual raters and refine the coarse prediction to form the final calibrated segmentation maps.
這是通過使用專業知識推斷模塊將評分者的專業知識線索明確整合到高級語義特徵中，以及使用多評分者協議建模模塊來重建單個評分者的評分並細化粗略預測以形成 最終校準的分割圖。

Extensive empirical experiments demonstrated the overall superior performance of our MRNet on a range of medical image segmentation tasks over diverse image modalities.
大量的經驗實驗證明了我們的 MRNet 在不同圖像模態的一系列醫學圖像分割任務上的整體優越性能。

Acknowledgement. 感謝

This work was funded by the Key-Area Research and Development Program of Guangdong Province, China (No. 2018B010111001), National Key R&D Program of China (2018YFC2000702) and the Scientific and Technical Innovation 2030-‘New Generation Artificial Intelligence’ Project (No. 2020AAA0104100), University of Alberta Start-up Grant, and NSERC Discovery Grants (No. RGPIN-2019-04575).
這項工作得到了廣東省重點領域研究發展計劃（No. 2018B010111001）、國家重點研發計劃

（2018YFC2000702）和科技創新 2030-“新一代人工智能”項目（No. . 2020AAA0104100)、阿爾伯塔大學啟動資助和 NSERC 發現資助（編號：RGPIN-2019-04575）。