Graph-based Region and Boundary Aggregation for Biomedical Image Segmentation

Yanda Meng, Hongrun Zhang, Yitian Zhao, Xiaoyun Yang, Yihong Qiao, Ian J. C. MacCormick, Xiaowei Huang, and Yalin Zheng

***Abstract*— Segmentation is a fundamental task in biomedical image analysis. Unlike the existing region-based dense pixel classification methods or boundary-based polygon regression methods, we build a novel graph neural network (*GNN*) based deep learning framework with multiple graph reasoning modules to explicitly leverage both region and boundary features in an end-to-end manner. The mechanism extracts discriminative region and boundary features, referred to as initialized region and boundary node embeddings, using a proposed Attention Enhancement Module (*AEM*). The weighted links between cross-domain nodes (region and boundary feature domains) in each graph are defined in a data-dependent way, which retains both global and local cross-node relationships. The iterative message aggregation and node update mechanism can enhance the interaction between each graph reasoning module's global semantic information and local spatial characteristics. Our model, in particular, is capable of concurrently addressing region and boundary feature reasoning and aggregation at several different feature levels due to the proposed multi-level feature node embeddings in different parallel graph reasoning modules. Experiments on two types of challenging datasets demonstrate that our method outperforms state-of-the-art approaches for segmentation of polyps in colonoscopy images and of the optic disc and optic cup in colour fundus images. The trained models will be made available at: `https://github.com/smallmax00/Graph_Region_Boudnary`**

***Index Terms*— Region-Boundary, Graph Neural Network, Segmentation**

I. Introduction

ACCURATE assessment of anatomical structures in medical images is critical in the management of a wide variety of medical conditions and diseases. For instance, glaucoma is a chronic neurodegenerative condition, and a leading cause of irreversible but preventable blindness worldwide [1]. The relative size of the optic disc (*OD*) and optic cup (*OC*) in colour fundus images is often used to assess glaucomatous damage to the optic nerve head [2], [3]. Similarly, colorectal polyps are positively associated with colorectal cancer, the third most common cancer worldwide [4]. Segmenting polyps provides essential information about the location and morphology of colorectal polyps for diagnosis and surgery. Manual annotation of these structures by clinicians is impractical because it is time-consuming, labour-intensive, and vulnerable to human error. Solving this problem depends on automated and precise biomedical image segmentation methods. To this end, we propose a graph-based deep learning framework to solve segmentation tasks, with the critical novelty of aggregating information about an object's region and boundary. We demonstrate the framework's effectiveness for segmentation of polyps in colonoscopy images and *OD & OC* in colour fundus images.

Previous deep learning based segmentation methods focused on learning the intensity features of the input image. They are either region-based methods performing dense pixel classification or boundary-based methods that regress the boundary's location. However, both neglect the intrinsic region-boundary relationship, which is critical for enhancing segmentation performance [5], [6]. For example, region features emphasise global homogeneity of pixel-wise semantics and object-level contextual information. On the other hand, boundary features describe the local edge characteristics and spatial variations on both sides of the boundary contour. Intuitively, combining information about region and boundary features ought to improve segmentation. Additionally, the subjective experience of clinicians who annotate biomedical images often involves assessing details of the relevant area as well as the boundary defining its margin. This may be especially true of regions with low contrast edges such as the *OC*, and clinicians typically traverse the cupped area to determine the *OC* boundary [7].

This paper demonstrates how to rationally combine region and boundary features using a single graph-structure model. This takes advantage of the proposed Graph Neural Network (*GNN*) model's long-range information propagation and cross-domain feature update capabilities. The summary pipeline of our work is depicted in Fig. 1, please refer to Fig. 2 for more details. The term 'cross-domain features' refers to the region features (containing semantic information) and

Y. Meng, H. Zhang and Y. Zheng are with the Institute of Life Course and Medical Sciences, University of Liverpool, Liverpool, L7 8TX, United Kingdom.

Y. Zhao is with the Cixi Institute of Biomedical Engineering, Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, Ningbo, 315201, China.

Y. Qiao is with the China Science IntelliCloud Technology Co., Ltd, Shanghai, China.

X. Yang is with Remark AI UK Limited, London, SE1 9PD, United Kingdom.

I.J.C. MacCormick is with the Centre for Inflammation Research, Queen's Medical Research Institute, University of Edinburgh, Edinburgh, EH16 4TJ, United Kingdom.

X. Huang is with the Department of Computer Science, University of Liverpool, Liverpool, L69 3BX, United Kingdom.

Corresponding author: Yitian Zhao (yitian.zhao@nimte.ac.cn); Yalin Zheng (yalin.zheng@liverpool.ac.uk).

boundary features (containing spatial information).

This paper explicitly considers information from both the region and boundary domains of objects of interest in medical images. Specifically, we construct multiple graphs, each contributing to tackling specific-level cross-domain feature updating and reasoning. Every graph contains region nodes with global semantic information and boundary nodes with local spatial characteristics. Weighted links between nodes exchange and aggregate semantic and spatial information. Additionally, we introduce an attention enhancement module (*AEM*) in conjunction with two sequential attention mechanisms through the channel and the spatial inter-dependencies. The *AEM* is built between the multi-level backbone features and the corresponding constructed graph nodes to extract discriminative feature embeddings for the region and boundary nodes, respectively. Further, we derive a spatial gradient from the predicted region mask as the derived boundary probability map to exploit the underlying consistency between the region and the boundary segmentation predictions. The discrepancies between the derived boundary probability map and the boundary ground-truth are formulated as one of the loss terms, dubbed boundary agreement loss, to enforce boundary consistency in region mask prediction during model training. Our experimental results show that the proposed *GNN*-based framework makes a significant improvement over the state-of-the-art methods.

In summary, this work makes the following contributions:

- The underlying relationship between the region and boundary characteristics is usually overlooked by approaches to segmenting biomedical images, despite human graders' instinctive use of both domains. We propose a novel end-to-end trainable segmentation model that integrates region and boundary features as graph nodes and updates and propagates cross-domain features.
- Cross-domain features are challenging to optimize concurrently; in particular, the inevitable prediction perturbation will impair joint cross-domain feature learning and updating. Here we introduce a boundary agreement loss function, ensuring that the predicted region and boundary mask have consistent boundaries.
- Extensive experiments demonstrate that our proposed model outperforms the state-of-the-art approaches on two segmentation tasks. Instead of conducting experiments on a small number of datasets, we combine five different *OD & OC* segmentation datasets and five different colonoscopy polyp segmentation datasets, respectively. In terms of varying dataset sources, they may contain different annotation standards for ground truths by various clinicians. Nevertheless, our model achieves good segmentation performance, demonstrating its robustness and generalizability.

II. Related Works

A. Region-based Segmentation

Convolutional Neural Networks (*CNN*) have found widespread applications in medical image segmentation. Existing CNN-based methods [8]–[16] have considered segmentation as a dense pixel classification task. For example, the classic *U-net* [12] employs a skip-connection between the encoder and decoder to alleviate information loss; and it has served as a baseline model for segmentation tasks in recent years. Another classic region-based segmentation method, *U-Net++* [9], uses an aggregated mechanism to fuse multi-level features. However, it may result in excessive information flow because some low-level features are unnecessarily over-extracted while object boundaries are simultaneously under-sampled. Recently, *Gu et al.* proposed *CE-Net* [13] to capture high-level information and preserve spatial information based on *U-Net* [12]. However, due to the limited receptive field of standard *CNN*, dense atrous convolutions were incorporated [17], [18] to enlarge the receptive regions for long-range context reasoning. *M-Net* [8] represented the fundus image in polar coordinates, and achieved high accuracy in segmenting *OD & OC*. However, it needed additional processes, such as multi-scale input and side-output mechanisms with deep supervision, to achieve multi-level receptive field fusion for long-range relationship aggregation. Similarly, *Fan et al.* proposed a *Inf-Net* [14] to tackle *COVID-19* lung infection segmentation. A reverse attention module is included to work with deep supervision in terms of multiple side-outputs. The aforementioned methods have achieved promising results in segmentation tasks with the help of boosted long-range relationship reasoning abilities. However, they are not efficient since stacking local cues cannot always precisely handle long-range context relationships. Especially for pixel-level classification problems, such as segmentation, performing long-range interactions is important for reasoning in complex scenarios [18]. To address this challenge, recent self-attention [19] based methods [15], [16] have demonstrated a superior ability to capture long-range relationships. For example, *Segtran* [16] proposed a squeezed attention block, which regularized the self-attention of *Transformers* [20], and an expansion attention block learned diversified representations. In this way, *Segtran* can calculate the pairwise interactions (self-attention) between all input units, combine their features and generate contextualized features. It has achieved promising results in the *OD & OC* and polyp segmentation tasks. On the other hand, in order to comprehend scenes or global contexts, these approaches must learn the object's position, boundary, and category from high-level semantic awareness and regional location information [21]. However, they tend to focus on learning image intensity features and suffer from a lack of regional position information at the pixel level [22]. This has resulted in inaccurate object boundary predictions.

B. Boundary-based Segmentation

Polygon-based boundary regression methods have drawn much recent attention. Polygon-based methods [23]–[27] regress the predefined vertex positions along the object boundaries and connect the predicted vertices to form a polygon, which is then converted into a mask. For example, *Cheng et al.* combined Active Contour Models (*ACMs*) [28] and *CNN*, to create a Deep Active Ray Network [23], which utilizes polar

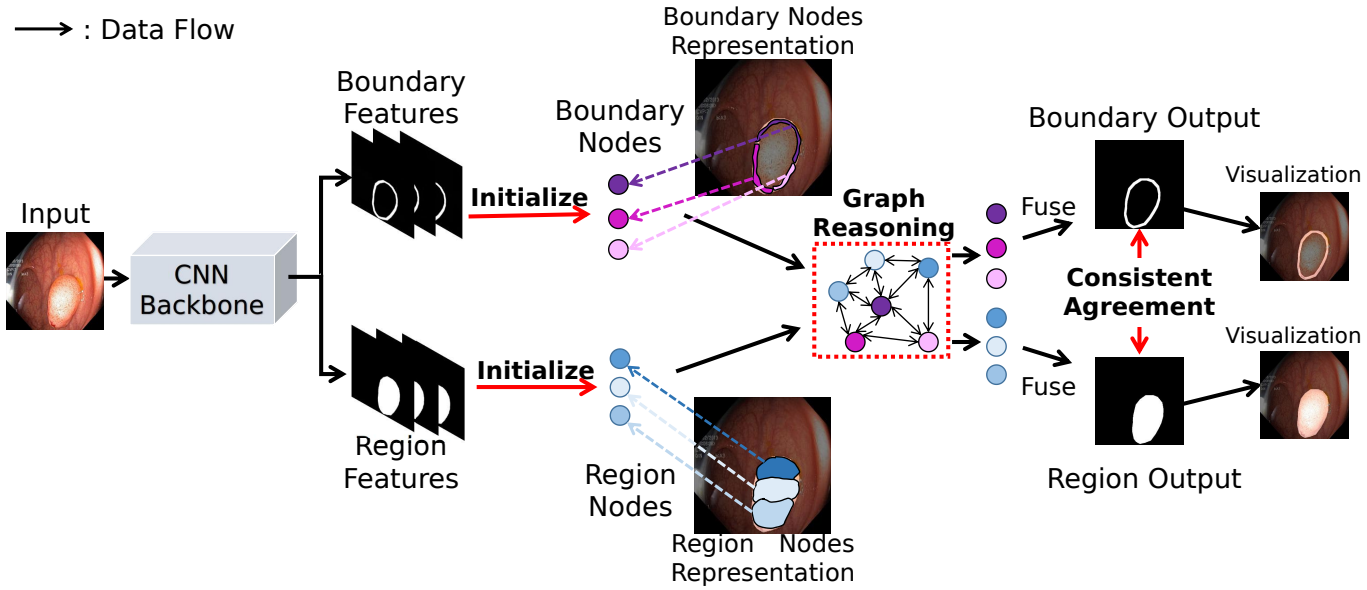

Fig. 1. The pipeline of the proposed network, with the example of a colonoscopy polyp image as the input. The extracted region and boundary features from the *CNN* backbone are treated as the initialized graph nodes and then go through the graph-level feature aggregation and reasoning process. A requirement for consistency between the boundary and the region outputs forces the *GNN* to learn coherent features.

coordinates (*rays*) to represent active contours. Along the same lines, *Xie et al.* proposed *PolarMask* [24] to interpret the object boundary in a polar coordinate system and proposed a *CNN* to regress the length of *rays*, which implicitly estimates the object boundary. Similarly, *Meng et al.* proposed *CABNet* [27], which represents the object boundary as vertices, then explicitly estimates the vertex locations. It achieved promising results on *OD & OC* segmentation tasks. Other boundary-based methods [6], [29]–[31] integrate the boundary geometry constraint into the loss function or evaluation measurement. For example, *Kervadec et al.* proposed boundary loss [31] which takes the distance metric on contours' space to mitigate the difficulties of highly unbalanced foreground and background. *Cheng et al.* proposed a Boundary Intersection-over-Union (*BIoU*) [6] evaluation measurement, which quantifies boundary quality in region segmentation tasks.

These methods are applicable to segment the whole region of the objects by regressing the position of vertices along boundary contours. However, they overlook the intrinsic region-boundary relationship, which we suggest is crucial for enhancing segmentation performance.

C. Region and Boundary for Segmentation

Recent methods, such as [5], [14], [32]–[37], explicitly or implicitly considered the dependency between the regions and boundaries of an object of interest in *OD & OC* or polyp segmentation. Specifically, *Zhang et al.* proposed *ET-Net* [32] for *OD & OC* segmentation, where an edge attention mechanism is proposed to explicitly emphasise the object boundary. On the other hand, *Fan et al.* [5], [14] and *Zhang et al.* [37] shared a similar boundary attention idea, where the object boundary is implicitly extracted from region predictions with a foreground erasing mechanism. In general these approaches treat segmentation as a multi-task learning problem, by using a shared backbone and two independent sub-networks to extract features of the regions and the boundaries, respectively. Then, the extracted features of regions and boundaries are directly fused with basic fusion operations such as element-wise addition or multiplication [5], [34], [37], or channel-wise concatenation [32], [36] with or without a fusion operation [33], [35].

We suggest that the correlations between region and boundary features cannot be adequately captured and exploited by two *independent* sub-networks that rely on these types of primary fusion operations. An intuitive solution would be to aggregate region and boundary features during the *whole learning process*. Unfortunately, the extracted region and boundary features are necessarily from two different domains and so contain varying semantic and spatial details. For example, region features focus on global homogeneity in pixel-wise semantics and object-level contextual information; while boundary features describe local edge characteristics and spatial variations on both sides of the boundary contours. It is well known that concurrently optimizing cross-domain features are difficult. Our experimental results also support this, and readers are directed to *Ablation Study* (Section V-A) for detailed information. In contrast, our method studies the cross-domain relationship of the region and boundary features throughout the whole training process with the help of the proposed *GNN* module. In other words, our model benefits from complementary cross-domain feature exchange and self-domain information propagation of region and boundary features along the entire training pipeline through the proposed graph structure model. Our experimental results prove that the proposed GNN reasoning module can tackle cross-domain feature optimization and achieved promising results on two segmentation tasks.

D. GNN in Segmentation

Graph-structure models have recently been adopted for segmentation tasks because of their natural aptitude for long-range information propagation and feature updates. *Dong et al.* [38] and *Shen et al.* [39] exploited the traditional random walk algorithm on a graph to tackle image segmentation tasks. However, the energy formulations for describing the images are complicated and higher-order energy function based methods [40], [41] may be needed to solve the problem. Recently, *Yao et al.* proposed a *GNN* network [42] to study the 3D geometrical relationship between vertices through mesh representation in an organ segmentation task. With the nature of *GNN*, long-range shape information can be updated and passed among vertices to maintain a consistency constraint. Along the same lines, *Voxel2mesh* [43] learned a deformable mesh representation through *GNN* to propagate the voxel features along the edges of the built graph model. Another paper [44] by *Shin et al.* used *GNN* to learn the global structure of the vessel's shape, which mirrored the connectivity of neighbouring vertices. Similarly, *Meng et al.* proposed *RBA-Net* [45] to regress the *OD & OC* boundaries by aggregated *CNN* and *GCN*, which learns the long-range features and directly regresses vertex coordinates in a Cartesian system.

The methods mentioned above used *GNN* to address the problem of intra-domain long-range feature propagation, as messages passing between graph nodes share similar semantic and spatial characteristics. In contrast, our method considers extracted region and boundary features as distinct graph nodes and employs *GNN* to learn their inter-domain relationship. Additionally, methods such as [42], [43], [45] represented each graph node with a predefined vertex and the corresponding coordinate under the form of mesh [42], [43] or triangle [45]. In that kind of framework, each graph node can only represent a single location. In contrast, our method represents each graph node with a set of pixels (locations) in the region area or boundary area (shown in Fig. 1).

III. Methods

Fig. 2 shows the model architecture of the proposed method. Given an input image, we extract the multi-level features through a backbone network. Following *PraNet* [5], we adopt the truncated *Res2Net* [46] as the backbone due to its superior ability to extract features in the segmentation task. We propose to use several *GNN* modules to reason and aggregate the extracted multi-level region and boundary features, which are elaborated as follows.

A. Attention Enhancement Module

Inspired by [47], we applied an attention enhancement module (*AEM*) upon each of the extracted multi-level backbone features. Specifically, the *AEM* is designed as a sequential operation consisting of channel attention $\mathbf{C}_{att}(\cdot)$ and spatial attention $\mathbf{S}_{att}(\cdot)$. The *AEM* is defined as: $F_{AEM}(f) = \mathbf{S}_{att}\left(\mathbf{C}_{att}(f)\right)$, where $\mathbf{C}_{att}(f) = f \otimes MLP(\mathbf{Pool}_c(f))$, $MLP(\cdot)$ is a multi-layer perceptron with two layers and sigmoid as the activation function; f is the input feature; $\mathbf{Pool}_c(\cdot)$ denotes the global max pooling for each feature map; $\otimes$ represents the multiplication by the dimension broadcast. In addition, $\mathbf{S}_{att}(f) = f \otimes Conv(\mathbf{Pool}_s(f))$, where $Conv(\cdot)$ is a 3×3 convolution layer with padding=1, followed by a sigmoid activation function; $\mathbf{Pool}_s(\cdot)$ denotes the global max pooling operation for each position in the feature map along the channel axis. In contrast to [47], we omitted the additional feature merging operations, such as the average pooling layer, in order to retain the most critical extracted characteristics.

As shown in Fig. 2, for each resolution's backbone feature map, we applied two *AEMs*, resulting in attention-enhanced region and boundary feature maps, respectively, which is referred to as the initialised nodes (region nodes $\mathbf{V}_r$ and boundary nodes $\mathbf{V}_b$). Fig. 1 demonstrates the boundary node and region node representations. Each node represents a set of relative features (pixels), such as region pixels and boundary pixels. The subsequent graph reasoning module treats each region and boundary nodes independently; afterwards, the output nodes of region and boundary are fused separately, resulting in region output $\mathbf{R}_p$ and boundary output $\mathbf{B}_p$. The whole network is end-to-end trainable; the supervision gradients of the region and boundary ground truth will back-propagate to the corresponding *AEM*, respectively. Thus, the two *AEM* will excavate the discriminative feature embeddings for the region and boundary features from each resolution's backbone feature.

B. Graph Based Reasoning

Fig. 2 illustrates several graphs in parallel that address the cross-domain, cross-level reasoning with varying numbers of region nodes $\mathbf{V}_r$ and boundary nodes $\mathbf{V}_b$. In this manner, the deep-level semantics of a region of interest, and the shallow-level spatial characteristics of the associated boundary can be interpreted as a whole. In the *Ablation Study* section, we perform detailed studies to evaluate the effectiveness of the number of graphs and the number of node updating times in each graph.

1) Graph Node Initialization: In our graph-based reasoning module, we construct multiple graphs in parallel, in which various levels of the attention-enhanced features are referred to as the initialized region node embeddings $\mathbf{V}_r = \{v_{r^1}, ..., v_{r^n}\}$ and boundary node embeddings $\mathbf{V}_b = \{v_{b^1}, ..., v_{b^n}\}$. In other words, we treat the extracted region and boundary output features of the *AEM* module as the corresponding region and boundary nodes in the proposed graph. The underlying motivations are twofold: (1) As mentioned before, the region and boundary output features from *AEM* contain different levels (shallow and deep) and domains (region and boundary) of information. In order to obtain complementary information from those features we treated them as graph nodes and used the message passing and information exchange mechanism of *GNN*. (2) In general, a GNN model propagates messages through a graph, with each node's representation conditioned on its relationships with surrounding nodes as well as its own information. Thus, through passing messages among different nodes, relevant information and relations may be gradually distilled for learning feature embeddings, where the region and boundary segmentation can be derived.

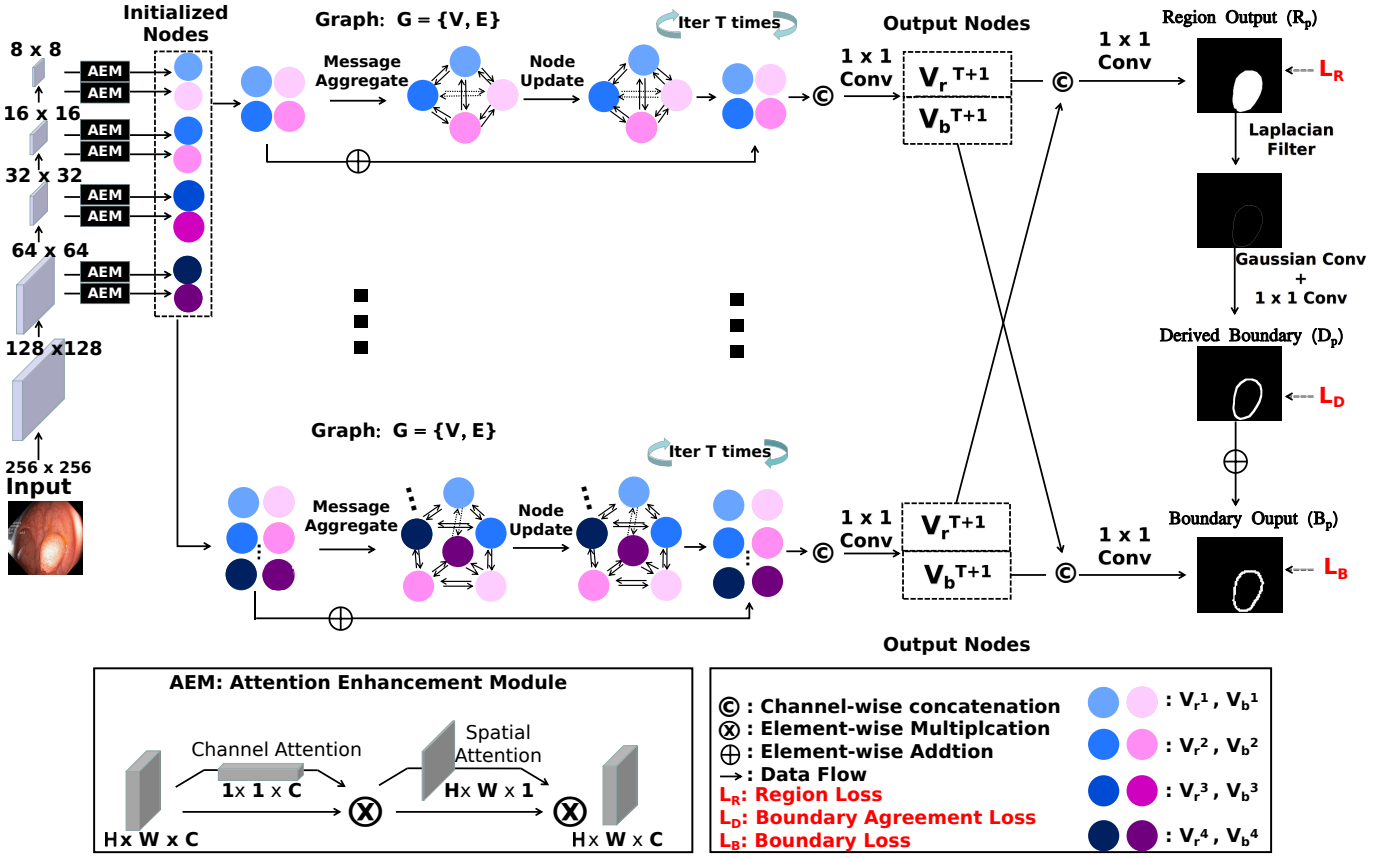

Fig. 2. Overview of the proposed *GNN* model (best viewed in color). The initialized nodes from the *AEM* output are interpolated into the same scale (32 × 32) through the bi-linear interpolation layer. For simplicity, we present only two graph reasoning modules in the middle, with the top one containing two region nodes and two boundary nodes from relatively deep feature level and the bottom one containing four region nodes and four boundary nodes from both shallow and deep feature levels. In this figure, we demonstrate how to segment polyps. As for *OD & OC* segmentation, the only difference is that the output probability map has a channel size of 2.

2) Single Graph Reasoning Module: In this section, we demonstrate the structure and components of a single graph, such as the one on the top middle in Fig. 2, in which there are four nodes with low-resolution (8×8 and 16×16); the one on the bottom middle has eight nodes of both low- and high-resolutions (from 8×8 to 64×64). Please note that, rather than being chosen at random, the nodes in each graph are fixed during training. Thus, each graph will address specific levels of the region and boundary feature aggregation process.

Node Embeddings. Given the initialized region nodes $\mathbf{V}_r = \{v_{r^1}, ..., v_{r^n}\}$ and boundary nodes $\mathbf{V}_b = \{v_{b^1}, ..., v_{b^n}\}$, we interpolate them to have the same size through the bi-linear interpolation layer. Then, we construct the graph $G = \{\mathbf{V}, \mathbf{E}\}$, where $\mathbf{V} = \mathbf{V}_r \cup \mathbf{V}_b$, are the combination of region and boundary nodes.

Edge Embeddings. For information propagation, nodes are linked with each other by weighted edges $\mathbf{E} = \{e_1, ..., e_{n^2-n}\}$, where the weighted edges can reflect the different correlations among various nodes. Rather than randomly initialising the edges, we define the edges in a data-dependent way. Inspired by [48], [49], for two linked nodes v_i, v_j from $\mathbf{V}$, the edge $\mathbf{e}_{i,j}$ from v_i to v_j is defined as:

$$\mathbf{e}_{i,j} = Conv(Cat(v_i - v_j, v_j)), \tag{1}$$

where $Cat(\cdot)$ is channel-wise concatenation, $Conv(\cdot)$ represents a 1×1 convolution layer to learn the relationships and minimise the channel size into 1. Thus, data-dependent local information $v_i - v_j$ and global information v_j are both considered in the edge $\mathbf{e}_{i,j}$. Note that, $e_{i,j}$ has the same size as v_i and v_j. In contrast, the edge $\mathbf{e}_{j,i}$ from v_j to v_i is defined as:

$$\mathbf{e}_{j,i} = Conv(Cat(v_j - v_i, v_i)). \tag{2}$$

In this way, the weighted edge embeddings contain the self-information of the starting node and the cross-information (cross domains or cross levels) of the connected node. Thus, both types of information can be aggregated to other connected nodes during the messaging passing process. The edge is defined as directional so as to distinguish the directional information passing and message aggregation among different nodes.

Message Aggregation & Nodes Update. In our *GNN* model, nodes connect with each other; as a result, each node aggregates the cross-level (deep and shallow) and cross-domain (region and boundary) messages from all its neighbouring nodes, then the node embeddings will be updated. At *T-th* update step, for the node v_i^{T-1} and all its neighbour nodes v_j^{T-1}, the message aggregation function $m_{j,i}^T$ from v_j^{T-1} to

v_i^{T-1} is defined as:

$$m_{j,i}^{T} = \sum_{j}^{n-1} ReLU(e_{j,i}^{T-1}) \odot v_j^{T-1}, \tag{3}$$

where $\odot$ is element-wise multiplication; $ReLU(\cdot)$ as the non-linear function to convert the edge embeddings to link weight. Then we update the node embeddings with a residual connection:

$$v_i^{T} = (\sum_{j}^{n-1} m_{j,i}^{T}) + v_i^{T-1}, \tag{4}$$

where the last step node embeddings v_i^{T-1} is maintained for the subsequent graph reasoning process.

After T times message aggregations and node updates, we fuse the region nodes $\mathbf{V}_r^{T+1} = \{v_{r^1}^{T+1}, ..., v_{r^n}^{T+1}\}$ and boundary nodes $\mathbf{V}_b^{T+1} = \{v_{b^1}^{T+1}, ..., v_{b^n}^{T+1}\}$ respectively through channel-wise concatenation, following by 1×1 convolution to generate the output region nodes and boundary nodes. T is 3 in our work.

3) Multi-level Graph Reasoning Modules: As observed by others [5], [9], the deep- and shallow- layer features from different levels complement one another, with the deep-layer features containing extensive semantic region information and the shallow-layer features retaining adequate spatial boundary information. To this end, we expand the proposed *GNN* by running several graph reasoning modules concurrently (2 in our work). Each graph includes region and boundary nodes from different shallow and deep feature levels of the backbone network. Thus, each graph reasoning module will address specific levels of aggregation and reasoning about region and boundary features. For example in Fig. 2, the top reasoning graph tackles the deep-level feature aggregation (8×8, 16×16), and the bottom reasoning graph tackles the shallow- and deep-level feature aggregation (8×8, 16×16, 32×32, 64×64). Finally, we fuse the output region ($\mathbf{V}_r^{T+1}$) and boundary nodes ($\mathbf{V}_b^{T+1}$) of each parallel graph respectively by channel-wise concatenation, followed by a 1×1 convolution with sigmoid activation function, then up-sample to obtain the region and boundary segmentation predictions ($\mathbf{R}_p$ and $\mathbf{B}_p$, with the same size of 256×256 as the input images). Please note, the parallel graphs are not connected during the reasoning process but have connections (fusion) on the output nodes. This is because each graph is designed to concentrate exclusively on a particular set of levels (resolutions) of nodal reasoning. We found that adding connections between graphs did not improve segmentation, but did increase training time.

C. Loss Function

The total loss function is defined as:

$$L_{total} = \mathbf{L}_R + \beta \cdot (\mathbf{L}_B + \mathbf{L}_D), \tag{5}$$

where **Dice Loss** [50] ($\mathbf{L}_R$) is used for the region segmentation predictions to penalize the mismatch regions against the corresponding ground truth. We defined L_R as:

$$L_R(R_p, Y_R) = 1 - \frac{2R_p GT_R + 1}{R_p + GT_R + 1}, \tag{6}$$

where R_p and GT_R denote the region segmentation predictions and the ground truth. Here, 1 is added to avoid divide by zero errors, such as when $R_p = GT_R = 0$. We also adopt the signed distance map loss (L_{sdm}) [31] as the boundary loss ($\mathbf{L}_B$) on boundary segmentation predictions due to the challenge of highly imbalanced foreground and background [51]. In detail, [31] used an integral approach for computing boundary variations with a signed distance transformation map, which can avoid complex local differential computations. Formally, the signed distance function (*SDF*) of segmentation ground truth (GT) can be defined as:

$$GT_{SDF} = \begin{cases} \underset{y \in \Delta G}{-inf} ||x - y||_2, & x \in GT_{in} \\ 0, & x \in \Delta G \\ \underset{y \in \Delta G}{inf} ||x - y||_2, & x \in GT_{out} \end{cases}$$

where $||x - y||_2$ represent the Euclidean distance between pixel x and y. Besides, GT_{out}, GT_{in} and ΔG, denote the outside, inside and boundary of the object, respectively. Given the signed distance maps of ground truth (GT_{SDF}) and the sigmoid outputs of the model $Pred_\theta$ (θ is the parameters), the signed distance map loss (L_{sdm}) is represented as:

$$L_{sdm}(Pred_\theta, GT_{SDF}) = Pred_\theta \odot GT_{SDF}, \tag{7}$$

where $\odot$ denotes Hadamard product. In this way, we can represent the boundary loss L_B in this work as:

$$L_B = L_{sdm}(B_p, GT_B), \tag{8}$$

where GT_B represents the signed distance map of the boundary segmentation ground truth. β is empirically set as 0.5 to balance the losses between Dice loss, region and boundary predictions.

Boundary Agreement Loss ($\mathbf{L}_D$). Firstly, we derive the spatial gradient from the predicted region mask (R_p), as the derived boundary probability map (D_p). In detail, we empirically adopt the *Laplacian* filter as a 3×3 kernel $[[1, 1, 1], [1, -8, 1], [1, 1, 1]]$ convolution layer to compute the spatial gradient in an end-to-end manner. The *Laplacian* filter is the direct result of a finite-difference approximation of the spatial derivative [52], highlighting the rapid intensity change regions. However, this will lead to thin and coarse derived boundaries, which results in extremely unlabeled classes (Shown in Fig. 2). To address this issue, we then empirically applied an approximated 3×3 *Gaussian* kernel convolution layer (*sigma* equals to 3 for two directions), followed by a 1×1 convolution layer to increase the boundary width and address the unbalanced issues [6]. The derived boundary probability map (D_p) is defined as:

$$D_p = Conv_{1\times1}\Big(Gaussian_{3\times3}\big(Laplacian_{3\times3}(R_p)\big)\Big). \tag{9}$$

Furthermore, the signed distance map loss [31] is applied to it against the boundary ground truth due to address the challenge of unbalanced classes.

The boundary agreement loss (L_D) is defined as:

$$L_D = L_{sdm}(D_p, GT_B). \tag{10}$$

With boundary agreement loss, region segmentation can benefit from additional boundary constraints, resulting in more reliable region segmentation predictions with more accurate boundary details. The boundary ground truth was generated by applying the same *Laplacian* filter and *Gaussian* kernel convolution to the corresponding segmentation ground truth mask. We then converted it into a binary map with threshold 0 as the final ground truth.

Furthermore, we empirically found that incorporating the derived boundary (D_p) into the boundary output (B_p) can enhance both the region and boundary segmentation performance. Thus, to augment the segmentation accuracy, we fuse the derived boundary probability map $\boldsymbol{D}_p$ with the boundary segmentation map $\boldsymbol{B}_p$ in terms of element-wise addition. The resulting concatenated feature map is then fed into a 1×1 convolution layer with a sigmoid activation function to produce the final boundary segmentation prediction. In this way, the boundary segmentation prediction $\boldsymbol{B}_p$ can benefit from the feature supplement provided by the derived boundary maps $\boldsymbol{D}_p$.

IV. Experiments

A. Datasets

We evaluate our approach with two distinct yet challenging medical image segmentation tasks: segmentation of *OD & OC* from retinal images, segmentation of polyps from colonoscopy images. Accurate segmentation of the *OC* in colour fundus images is often difficult because of poor contrast between the cup and the surrounding rim [7]. The boundary between a polyp and its surrounding mucosa is typically blurred in colonoscopy images and lacks the intense contrast required for segmentation approaches [53].

Fundus images of OD and OC: We pooled 2068 images from five datasets (Refuge [7], Drishti-GS [54], ORIGA [55], RIGA [56], RIM-ONE [57]). 613 fundus images were randomly selected as the test dataset, leaving the other 1455 images for training and validation. Following [45], we located the disc center from each image and then cropped a subimage of 256 $\times$ 256 pixels centered on the disc for the subsequent analysis.

Colonoscopy polyp images: We retrieved 2085 colonoscopy images from five datasets (ETIS [4], CVC-ClinicDB [58], CVC-ColonDB [59], EndoScene-CVC300 [60], and Kvasir [53]). We used the same data split settings as [5], namely 1450 colonoscopy images from Kvasir [53] and CVC-ClinicDB [58] comprised the training and validation datasets. The remaining 635 colonoscopy images from [4], [59], [60] were used for testing. All of the images are uniformly resized to 256 $\times$ 256.

B. Experimental Setting and Evaluation Metrics

To augment the dataset, we randomly rotated and horizontally flipped the training dataset with a probability of 0.3. The rotation ranges from -30 to 30 degree. We use stochastic gradient descent with a momentum of 0.9 to optimize the overall parameters. We trained the model around 300 epochs for all the experiments, with a learning rate of 1e-2 and a decay rate of 0.5 every 100 epochs. The batch size was set as 48. The network was trained end-to-end; all the training processes were performed on a server with 4 TESLA V100, and all the test experiments were conducted on a local workstation with *Intel(R) Xeon(R) W-2104 CPU* and *Geforce RTX 2080Ti GPU* with 11GB memory. Five-fold cross-validation was used for fair comparison and hyper-parameters tuning in all settings. We randomly selected 10% of the training dataset for internal validation.

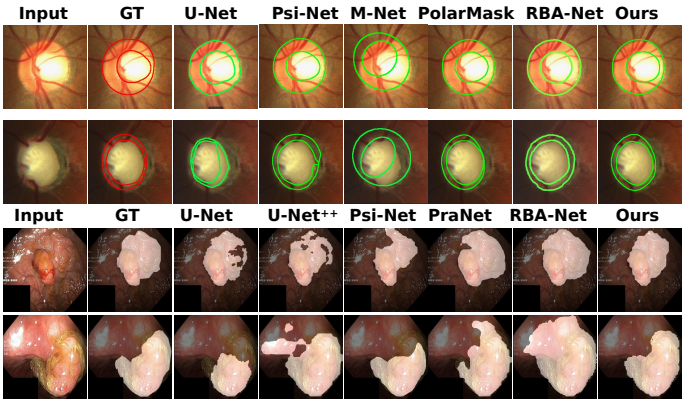

Fig. 3. Qualitative results of *OD & OC* segmentation and colonoscopy polyp segmentation. We compare our model with *U-Net* [12], *U-Net++* [9], *M-Net* [8], *PolarMask* [24], *PraNet* [5], *Psi-Net* [35], *RBA-Net* [45]. Our method can produce more accurate segmentation results when compared with ground truth (*GT*). Note that we plot the boundary (spatial gradient through *Laplacian* filter) of the region mask on the input image to better visualise the *OD & OC* segmentation comparison. Along the same lines, we highlight the region in the input image for colonoscopy polyp segmentation comparison.

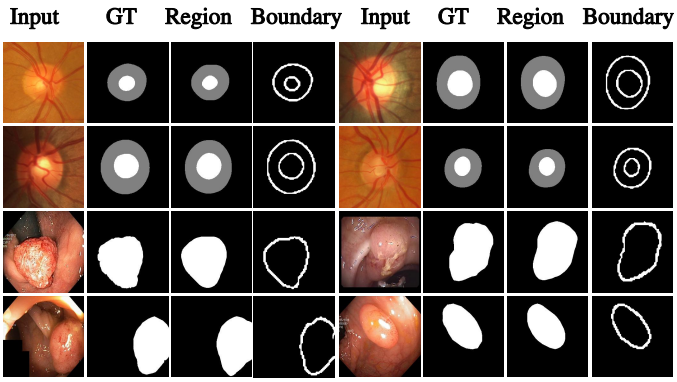

Fig. 4. Figure shows the binary mask comparison between our model's prediction and the ground truth. Our model produces consistent region **(Region)** and boundary **(Boundary)** predictions compared with the ground truth **(*GT*)**.

We report Dice similarity score (*Dice*) and balanced accuracy (*B-Acc*) as the region segmentation accuracy metrics; and Boundary Intersection-over-Union (*BIoU*) [6] as the boundary segmentation metric. 95% confidence intervals were generated by using 2000 sample bootstrapping. As for *BIoU* [6], compared with other boundary-based evaluation metrics such as *Trimap IoU* [18], [61] or *Boundary F1-measure* [62], [63], *BIoU* is more sensitive to show boundary errors on small objects (*e.g.* polyps) [6]. *BIoU* is defined as:

$$BIoU = \frac{|(B_p \cap Y_B) \cap (R_p \cap Y_R)|}{|(B_p \cap Y_B) \cup (R_p \cap Y_R)|}, \tag{11}$$

where Y_B and Y_R are the boundary segmentation ground truth

TABLE I

QUANTITATIVE SEGMENTATION RESULTS OF *OD & OC* AND POLYPS ON RESPECTIVE TESTING DATASETS. THE PERFORMANCE IS REPORTED AS *Dice* (%) AND *B-Acc* (%) AND *BIoU* (%). 95% CONFIDENCE INTERVALS ARE PRESENTED IN THE BRACKETS, RESPECTIVELY. WE COMPARE OUR MODEL WITH PREVIOUS STATE-OF-THE-ART METHODS BY RUNNING THEIR OPEN-SOURCE CODE. NOTABLY, WE SAMPLED 120 VERTICES FOR *PolarMask* [24], *CABNet* [27] AND *RBA-Net* [45] TO CONSTRUCT A SMOOTH BOUNDARY.

Tasks / Methods	OC			OD			Polyps		
	Dice (%)↑	*B-Acc (%)*↑	*BIoU(%)*↑	*Dice (%)*↑	*B-Acc (%)*↑	*BIoU(%)*↑	*Dice (%)*↑	*B-Acc (%)*↑	*BIoU(%)*↑
U-Net [12]	85.3 (82.1, 86.8)	87.1 (85.9, 88.8)	80.1 (77.6, 82.4)	95.0 (93.1, 97.1)	97.0 (95.3, 98.6)	86.2 (84.1, 88.3)	66.7 (63.6, 68.1)	73.7 (72.1, 75.1)	60.0 (57.6, 62.2)
U-Net++ [9]	86.0 (83.8, 88.5)	87.6 (85.3, 89.1)	81.4 (79.5, 83.8)	95.0 (93.9, 96.1)	97.9 (97.0, 98.5)	88.0 (86.4, 89.8)	65.6 (63.1, 67.7)	72.6 (70.1, 74.4)	58.8 (55.6, 61.3)
M-Net [8]	86.9 (85.0, 88.0)	89.7 (88.3, 90.9)	82.9 (79.5, 84.7)	96.8 (95.5, 97.6)	96.7 (95.9, 97.9)	88.1 (87.0, 89.3)	-	-	-
PolarMask [24]	87.2 (85.3, 89.1)	90.9 (88.7, 91.6)	83.2 (81.0, 85.1)	96.5 (95.8, 97.2)	97.8 (96.9, 98.5)	87.0 (86.0, 88.3)	69.3 (67.2, 71.4)	83.6 (81.2, 85.7)	60.3 (58.4, 61.9)
PraNet [5]	-	-	-	-	-	-	74.0 (72.6, 75.7)	85.6 (84.1, 86.9)	66.0 (63.3, 68.9)
Psi-Net [35]	85.7 (83.0, 88.2)	87.1 (85.5, 89.0)	82.1 (80.3, 84.0)	95.8 (94.5, 97.1)	97.7 (96.5, 98.4)	87.9 (85.4, 89.2)	63.8 (59.7, 65.9)	75.5 (73.1, 77.2)	57.1 (55.7, 58.6)
RBA-Net [45]	87.8 (85.2, 89.7)	89.5 (87.1, 91.6)	83.8 (81.6, 85.9)	96.1 (95.5, 96.7)	97.5 (96.4, 98.1)	88.9 (88.0, 89.2)	73.5 (71.2, 75.6)	85.1 (83.0, 87.3)	66.2 (64.8, 67.9)
ACSNet [37]	-	-	-	-	-	-	70.1 (67.8, 72.3)	82.6 (80.8, 84.4)	63.3 (60.1, 65.7)
CABNet [27]	87.1 (84.9, 88.8)	88.8 (87.1, 90.2)	83.0 (81.1, 85.4)	95.5 (94.6, 96.7)	96.4 (95.5, 97.2)	88.2 (87.1, 89.6)	73.0 (70.7, 75.4)	84.2 (82.0, 86.3)	65.5 (63.2, 67.7)
Segtran [16]	88.8 (86.5, 90.3)	91.0 (88.6, 93.2)	83.9 (81.3, 85.8)	97.3 (96.1, 98.2)	97.5 (96.6, 98.8)	90.0 (89.1, 91.2)	75.3 (73.5, 77.1)	86.5 (84.4, 88.3)	67.9 (65.5, 69.2)
Ours	**89.4** (87.6, 90.8)	**91.7** (91.1, 92.5)	**85.1** (83.3, 86.8)	**97.7** (97.0, 98.7)	**98.1** (97.8, 98.5)	**91.1** (90.2, 92.0)	**75.7** (73.1, 77.6)	**87.0** (86.1, 88.3)	**69.3** (67.9, 70.5)

TABLE II

ABLATION STUDY ON DIFFERENT FEATURE FUSION METHODS. THE PERFORMANCE IS REPORTED AS *Dice* (%), *BIoU* (%), ON THE TWO SEGMENTATION TEST DATASETS.

Tasks / Methods	OC		OD		Polyps	
	Dice (%)↑	*BIoU (%)*↑	*Dice (%)*↑	*BIoU (%)*↑	*Dice (%)*↑	*BIoU (%)*↑
w/o Fusion	86.6	79.1	94.7	86.7	71.2	64.6
w/ Addition	87.0	81.7	96.0	86.6	70.9	63.0
w/ Concatenation	85.7	80.1	94.8	87.5	71.1	65.3
w/ Non-local [19]	87.2	83.4	95.2	89.6	74.9	69.1
w/ *GloRe* [64]	88.1	84.3	96.1	89.9	73.7	67.5
Ours	**89.4**	**85.1**	**97.7**	**91.1**	**75.7**	**69.3**

and the region segmentation ground truth, respectively; R_p and B_p are the region and boundary predictions.

C. Performance Comparison and Analysis

In this section, we show qualitative (Fig. 3, Fig. 4) and quantitative (TABLE I) results of the *OD & OC* and polyp segmentation tasks. The best result in each category is highlighted in bold.

***OD & OC* Segmentation** Fig. 3 and 4 show qualitative results. TABLE I provides the quantitative results of *Ours* and other methods. We obtain an average 89.4% and 97.7% *Dice* on *OC* and *OD* segmentation, respectively, outperforming approaches based on region segmentation such as *U-Net++* [9] and *M-Net* [8] by an average of 3.4% and 1.9% respectively; outperforming polygon-based boundary regression approaches such as *PolarMask* [24] by 1.9%; outperforming boundary-region based methods such as *Psi-Net* [35] by 3.2%; and outperforming *GNN* based segmentation methods such as *RBA-Net* [45], *CABNet* [27] by 1.8% and 2.5%. Note that *PraNet* [5] and *ACSNet* [37] are specially designed for binary segmentation of colorectal polyps with respect to the implicit region-boundary reverse attention module. We cannot extend it to *OD & OC* segmentation directly since this is a multi-segmentation task. On the other hand, training two models, one for *OD* segmentation and another for *OC* segmentation, would be unfair to the other models under comparison. As a result, this model was not tested on the *OD & OC* segmentation tasks.

Polyp Segmentation TABLE I and Fig. 3, Fig. 4 show the quantitative and qualitative results. Our model achieves 75.7% *Dice*, which outperforms the cutting-edge *ACSNet* [37] and *PraNet* [5] by 8.0% and 2.2% respectively. As for boundary segmentation accuracy, our model achieves 69.3% *BIoU*, which is 5.0% better than *PraNet* [5] and 8.0% better than *ACSNet* [37]. Our model size ($\sim$ 38.69 million parameters) is larger than *PraNet* [5] ($\sim$ 30.49 million parameters) when our framework has 2 graph reasoning modules (shown in TABLE IV). However, our model can gain more accurate segmentation performance (74.3% *Dice*; 68.1% *BIoU*) with a comparable model size ($\sim$ 30.57 million parameters) with *PraNet* [5] when the number of graph reasoning modules is 1 (N = 1 in TABLE IV). *Segtran* [16] is a very recent region-based approach for polyp segmentation. It benefits from the long-range feature reasoning ability of *Transformer* [20], and achieves comparable performance with *ours*. However, it has a larger model size (93.0 million parameters) than *ours* (38.69 million parameters), and due to the complexity of the model structure it has a relatively lower inference speed (8.7 *fps*) compared with *ours* (21.6 *fps*) on our local machine).

V. DISCUSSION AND CONCLUSION

A. Ablation Study

We conducted detailed ablation studies, and all the results demonstrate our model's effectiveness. As an illustration, the ablation results for different feature fusion methods, network components, attributes of the graph reason modules, and loss functions are shown in TABLE II, TABLE III, TABLE IV, and TABLE V.

Feature Fusion. In this section, we evaluated the effectiveness of the proposed *GNN* reasoning module. Firstly, we replaced the *GNN* module with two feed-forward *CNN* blocks for the region and boundary features, respectively, to minimise the model size gap and retain a comparable number of parameters (e.g., $\sim$ 38.69 million for our model). In each *CNN* block, we built several standard convolution layers with kernel size 3 $\times$ 3, padding 1, followed by a Batch Normalization layer. Then, the boundary and region features are fused in three ways (similar to previous methods [5], [32]–[36]), including element-wise addition [5], [34], channel-wise concatenation [32], [36] or without fusion operation [33], [35]. Finally, two 1 $\times$1 convolution layers were added to generate the region and boundary predictions. Additionally, we adopted two more potent fusion mechanisms to show our proposed *GNN* reasoning module's superiority. In detail, we applied the Non-local module [19], and *GloRe* module [64] respectively, where the Non-local module exploits a self-attention mechanism [20] and *GloRe* utilizes graph convolution [65] to tackle the long-range relations among features. TABLE II shows that our model with the *GNN* reasoning module achieves much more accurate and reliable results than simple fusion operations and outperforms the Non-local and *GloRe* modules by 2.2% and 2.0% in terms of *Dice* (%); and 1.4% and 1.7% in terms of *BIoU* (%) on two segmentation tasks respectively.

TABLE III
ABLATION STUDY ON DIFFERENT MODEL STRUCTURE COMPONENTS. THE PERFORMANCE IS REPORTED AS *Dice* (%), *BIoU* (%), ON THE TWO SEGMENTATION TEST DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Methods \ Tasks	OC		OD		Polyps	
	Dice (%)↑	*BIoU* (%)↑	*Dice* (%)↑	*BIoU* (%)↑	*Dice* (%)↑	*BIoU* (%)↑
w/o *AEM*	87.1	83.6	96.7	89.2	74.0	68.0
w/o *Gaussian*	88.3	84.7	97.1	90.2	74.6	68.1
w/o Boundary nodes	86.3	82.0	94.8	88.7	72.9	66.0
w/o Region nodes	83.6	80.2	91.2	87.5	64.1	57.9
Ours	**89.4**	**85.1**	**97.7**	**91.1**	**75.7**	**69.3**

Network Components. This section presents the results of our ablation study on network structure components. We evaluated the effectiveness of the attention enhancement module (*AEM*), *Gaussian* kernel convolution layer, boundary nodes, and region nodes, respectively. We did this by removing each of those components in turn while retaining the rest of the structure. Notably, we overlooked the model size difference for the ablation study of the *AEM* and the *Gaussian* kernel convolution layer because there is no significant difference in the number of model parameters. To retain a comparable model size for the boundary nodes and region nodes ablation studies, we added feed-forward *CNN* blocks (same as the one in the Feature Fusion ablation study) after the *GNN* reasoning module. (1). The *AEM* is designed to extract the discriminating features for boundary and region nodes through the back-propagation mechanism in the proposed end-to-end trainable network. TABLE III demonstrates that our model (*Ours*) improves average 2.1% *Dice* and 2.0% *BIoU*, respectively, using *AEM* upon two segmentation test datasets.

(2). The *Gaussian* kernel convolution layer (*Gaussian*) is critical to increasing the boundary width in the generation of boundary ground truth and the derived boundary prediction (D_p). As discussed previously, we use it to increase the boundary width of the boundary output (B_p) and of the boundary ground truth. Our model (*Ours*) gains 1.1% *Dice* and 1.6% *BIoU* improvement upon two segmentation tasks.

(3). We performed extensive experiments to evaluate the significance of boundary nodes and region nodes by removing every element associated with the boundary nodes, including the corresponding *AEM*, V_b, D_p, B_p, L_D, L_B, *etc.*. In this way, the network is devoid of boundary information supervision and produces only region prediction. Furthermore, the proposed *GNN* module can only serve as a cross-level (shallow and deep) feature refinement module for the region segmentation task. It shows that our model (*Ours*) gains 3.5% *Dice* and 3.1% *BIoU* improvement from boundary information supervision on two segmentation tasks. On the other hand, we remove region information related elements in the network such as the corresponding *AEM*, V_r, D_p, R_p, L_D, L_R, *etc.*, and construct a boundary segmentation network. TABLE III shows that the model cannot achieve comparatively promising segmentation results due to the lack of supervision over region details. This further demonstrates the importance of boundary and region information in biomedical image segmentation tasks.

Attributes of the graph reason modules. In this section, we present the results of the ablation study on the attributes of the graph reason modules. Here we evaluated the effectiveness of the number of graph reasoning modules (N) and the number of update times (T) in each graph reasoning module. TABLE IV shows that our model achieves the best performance on two segmentation test datasets with two graph reasoning modules (N = 2), and each module updates three times (T = 3). In detail, the two graph reasoning model tackles (8 $\times$ 8, 16 $\times$ 16) and (8 $\times$ 8, 16 $\times$ 16, 32 $\times$ 32, 64 $\times$ 64) levels' features, respectively.

Furthermore, the number of graph reasoning modules (N) impacts the model size; the number of update times in each graph can influence inference time. To present a comprehensive analysis, we show the inference time and model size with different attributes of the graph reason module in TABLE IV. As shown, with N = 2 and T = 1, our model can run at a real-time speed of $\sim$ 38.1 *fps* and $\sim$ 44.0 *fps* for a 256 $\times$ 256 input of fundus image and colonoscopy image, respectively.

Loss Function. In general, the losses employed in this work serve a variety of purposes. Dice loss [50] ($\mathbf{L}_R$) is a commonly used region-based loss for segmentation task. While Dice loss outperforms other losses (i.e. Cross-Entropy loss) in addressing the unbalanced issues [31], we discover that by using Dice loss for boundary segmentation, the predicted boundary

TABLE IV

ABLATION STUDY ON THE ATTRIBUTES OF THE GRAPH REASON MODULES. THE SEGMENTATION PERFORMANCE IS REPORTED AS *Dice* (%), *BIoU* (%); THE INFERENCE SPEED IS REPORTED AS FRAME PER SECOND (*fps*) ON THE TWO TESTING DATASETS. ADDITIONALLY, WE PRESENT THE MODEL SIZE IN MILLIONS OF PARAMETERS. THE BEST RESULT IN EACH CATEGORY IS HIGHLIGHTED IN BOLD.

Methods \ Tasks	*OD & OC*			Polyps			*Model Size*
	Inference (fps) ↑	*Dice* (%)↑	*BIoU* (%)↑	*Inference (fps)*↑	*Dice* (%)↑	*BIoU* (%)↑	*(# of parameters in millions)*↓
$N = 1, T = 3$	∼21.6	92.1	86.6	∼29.3	74.3	68.1	∼30.57
$\mathbf{N = 2, T = 3}$	∼21.6	93.6	88.1	∼29.3	75.7	69.3	∼38.69
$N = 3, T = 3$	∼21.6	91.8	86.1	∼29.3	72.1	66.0	∼46.56
$N = 2, T = 1$	∼38.1	92.0	87.4	∼44.0	74.8	68.3	∼38.69
$\mathbf{N = 2, T = 3}$	∼21.6	93.6	88.1	∼29.3	75.7	69.3	∼38.69
$N = 2, T = 5$	∼3.7	91.9	87.3	∼13.8	73.4	68.1	∼38.69

TABLE V

ABLATION STUDY ON THE LOSS FUNCTION. THE PERFORMANCE IS REPORTED AS *Dice* (%), *BIoU* (%) ON TWO SEGMENTATION TEST DATASETS. THE BEST RESULT IN EACH CATEGORY IS HIGHLIGHTED IN BOLD.

Methods \ Tasks	OC		OD		Polyps	
	Dice (%)↑	*BIoU* (%)↑	*Dice* (%)↑	*BIoU* (%)↑	*Dice* (%)↑	*BIoU* (%)↑
w/ Dice Loss	87.0	83.2	95.2	89.0	73.3	67.0
w/o Agreement Loss	88.1	84.1	96.2	90.0	74.2	67.9
Ours	**89.4**	**85.1**	**97.7**	**91.1**	**75.7**	**69.3**

segmentation masks appear to be incomplete, leading to almost black masks (most zero pixel values) due to the unbalanced foreground and background. We addressed this challenge by applying boundary loss [31] ($\mathbf{L}_B$) to the boundary segmentation predictions (B_p). Boundary agreement loss ($\mathbf{L}_D$) adopts [31] as well. However, it is applied on the derived boundary (D_p), which aims for the consistent boundary upon the region predictions (R_p) and boundary predictions (B_p). L_D brings two essential advantages. Firstly, since D_p and B_p are under the supervision of the same boundary ground truth, L_D can be considered as the consistency loss between the D_p and B_p; at the same time, it can force the model to learn consistent boundary features for region nodes V_r and boundary nodes V_b. Secondly, the L_D serves as a boundary focus on the R_p with additional boundary ground truth supervision. This aids the model to produce more precise boundary predictions.

To analyse the effectiveness of the L_B and L_D, we applied Dice loss [50] to L_B (w/ Dice Loss), which is inevitably vulnerable to unbalanced foreground and background. TABLE V shows that our model improves by 2.9% *Dice* and 2.7% *BIoU* with boundary loss [31] on two segmentation tasks. Additionally, we excluded boundary agreement loss (L_D) while maintaining the remaining components to verify its importance (w/o Boundary Agreement Loss). As shown, L_D can deliver a 1.7% *Dice* improvement in region segmentation and 1.5% *BIoU* improvement in boundary segmentation.

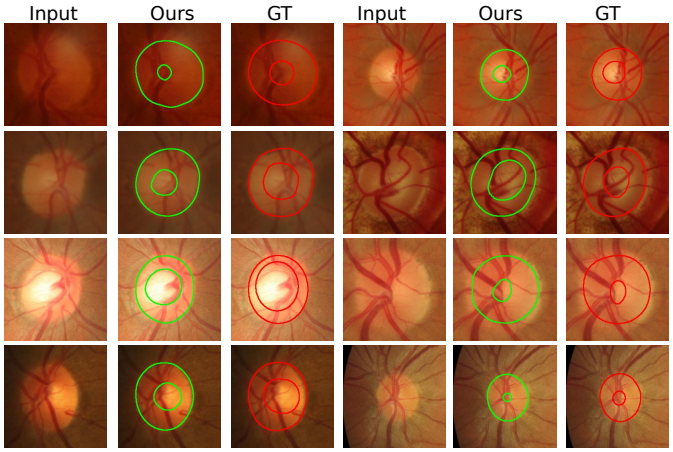

Fig. 5. A comparison of our segmentation (green) and the ground truth (red) in some 'failed' cases. The ground truth has inaccurate OC boundaries for most of the cases. According to an ophthalmologist (IJCM), our model generally produces more precise boundaries than the ground truth.

B. Clinical Evaluation and 'Failure' Analysis

Clinical Evaluation.

As well as assessing computer vision evaluation metrics, we also evaluated the clinical output of our method. The vertical Cup to Disc Ratio (*vCDR*) is an important indicator for screening and diagnosis of glaucoma. The *vCDR* value is calculated by the ratio of vertical cup diameter to vertical disc diameter. A larger *vCDR* indicates a higher possibility of glaucoma and vice versa. Following previous methods [8], [45], we provided the Mean Absolute Error of *vCDR* (δ_{vCDR}) between the predictions and the ground truth. Our method (*Ours*) achieved 0.056 δ_{vCDR} on the *OC & OD* segmentation test set, which outperformed classic methods *U-Net* [12] (0.089 δ_{vCDR}) and *U-Net++* [9] (0.077 δ_{vCDR}) by 37.1 and 27.3% respectively, outperformed cutting-edge methods *M-Net* [8] (0.064 δ_{vCDR}), *RBA-Net* [45] (0.062 δ_{vCDR}), *Segtran* [16] (0.060 δ_{vCDR}) and *CABNet* [27] (0.067 δ_{vCDR}) by 12.5%, 9.7%, 6.7% and 16.4%. *Ours* provides more accurate *vCDR* estimation than these other methods, and this is consistent with superior segmentation.

'Failure' Analysis. We studied the reasons for poor segmen-

tation by our method, and found that in some cases this could be attributed to imprecise ground truth in public *OD & OC* segmentation datasets. In detail, for each retinal image in the *OD & OC* test dataset, we considered segmentation to have 'failed' when the *Dice* (%) of *OC* segmentation was below 80.0% or *OD* segmentation was below 90.0%. According to these criteria, segmentation failed on 28 out of 613 test images. We made a montage of each case, comprising the original image, our segmentation, and ground truth. We present some of the failed segmentations by using our model (*Ours*) and the ground truth (*GT*) in Fig. 5. The ophthalmologist (IJCM) reviewed these 28 montages in a masked manner and indicated which of the two segmentations was more accurate for *OC* and *OD*, respectively. A McNemar-Bowker test [66] confirmed that *Our* segmentation was regarded as clinically accurate significantly more often than the *GT* (p=0.029 for *OC* and p=0.001 for *OD*). Further subjective clinical review of some *GT* imagesets suggested variable *GT* accuracy. This highlights the robustness of our model, but also points to important limitations in the ground truth manual annotations. The quality of manual annotations is of utmost importance for developing and validating segmentation models as well as translating automation tools into clinical practice. We advise investigators to apply extra caution when using public datasets. Quality assurance of manual annotations of public datasets is a strategic vulnerability in the field and requires further work.

C. Limitation and Future Work

Limitations

Our method achieves promising results for segmenting *OC & OD* and colonoscopy polyps. However, it may not work as well for highly complex objects, such as curvilinear structures like retinal vessels [30], [67], [68]. The primary reason for this is that retinal vessels' region and boundary areas can be challenging to distinguish due to their complex topology and tortuosity. In particular, the derived boundary map (D_p) we propose may have a significant overlap with the region map (R_p) in these situations. Thus, an inevitable perturbation will be included in the information propagation and message passing process between the region and boundary nodes, harming the segmentation performance.

Future Work.

Our method can be extended to tackle video-based segmentation tasks, especially for polyp segmentation. In brief, video-based polyp segmentation methods require high accuracy and speed at the same time. In addition, polyps are of varying size, and their appearance depends on the movement of the camera past the lesion. Thus, dynamic and rapid updates to the receptive field of the network are essential. An extension from our proposed multi-level graph reasoning modules, where each graph is responsible for tackling a specific level of the receptive field, a dynamic attention module (similar to [15]) could be applied on the fusion of different graphs. In this way, our model could automatically adopt the weight contributions between different graphs for inference predictions. As for the inference speed required by video-based tasks, a trade-off between accuracy and speed can be achieved by a different number of graphs and iteration numbers for message passing. Besides this, our proposed model could also be extended to tackle 3D image-based segmentation tasks. In 3D settings, we can regard the boundary as a surface mesh (vertices) and the region as voxels. Thus, the proposed boundary nodes in our method could represent the extracted surface mesh (vertices) features, and the region nodes could represent the extracted voxel-wise features. In this case information exchange and message passing between the surface and volume of 3D objects could be achieved with the same network, simply by redefining the identity of the nodes.

D. Conclusion

We propose a novel graph-based aggregation module that takes advantage of intuitive associations between the region and boundary features in biomedical images, in order to produce more accurate segmentation. Our experiments have demonstrated that the proposed model can effectively aggregate and explain the semantic region features and spatial boundary features for segmentation of polyps from colonoscopy images, and the optic disc & optic cup from retinal images. We believe the proposed *GNN* model can also tackle other cross-domain feature reasoning challenges, such as regions, boundaries, and landmark reasoning segmentation tasks.

REFERENCES

[1] Y.-C. Tham, X. Li, T. Y. Wong, H. A. Quigley, T. Aung, and C.-Y. Cheng, "Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis," *Ophthalmology*, vol. 121, no. 11, pp. 2081–2090, 2014.

[2] M. S. Haleem, L. Han, J. Van Hemert, and B. Li, "Automatic extraction of retinal features from colour retinal images for glaucoma diagnosis: a review," *Computerized Medical Imaging and Graphics*, vol. 37, no. 7-8, pp. 581–596, 2013.

[3] T. Li, W. Bo, C. Hu, H. Kang, H. Liu, K. Wang, and H. Fu, "Applications of deep learning in fundus images: A review," *Medical Image Analysis*, p. 101971, 2021.

[4] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer," *International Journal of Computer Assisted Radiology and Surgery*, vol. 9, no. 2, pp. 283–293, 2014.

[5] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Pranet: Parallel reverse attention network for polyp segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 263–273.

[6] B. Cheng, R. Girshick, P. Dollár, A. C. Berg, and A. Kirillov, "Boundary IoU: Improving object-centric image segmentation evaluation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[7] J. I. Orlando, H. Fu, J. B. Breda, K. van Keer, D. R. Bathula, A. Diaz-Pinto, R. Fang, P.-A. Heng, J. Kim, J. Lee *et al.*, "REFUGE challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs," *Medical Image Analysis*, vol. 59, p. 101570, 2020.

[8] H. Fu, J. Cheng, Y. Xu, D. W. K. Wong, J. Liu, and X. Cao, "Joint optic disc and cup segmentation based on multi-label deep network and polar transformation," *IEEE Transactions on Medical Imaging*, vol. 37, no. 7, pp. 1597–1605, 2018.

[9] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 3–11.

[10] L. Mou, Y. Zhao, L. Chen, J. Cheng, Z. Gu, H. Hao, H. Qi, Y. Zheng, A. Frangi, and J. Liu, "CS-Net: Channel and spatial attention network for curvilinear structure segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 721–730.

[11] L. Mou, Y. Zhao, H. Fu, Y. Liu, J. Cheng, Y. Zheng, P. Su, J. Yang, L. Chen, A. F. Frangi *et al.*, "CS2-Net: Deep learning segmentation of curvilinear structures in medical imaging," *Medical Image Analysis*, vol. 67, p. 101874, 2021.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 2015, pp. 234–241.

[13] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, and J. Liu, "CE-Net: Context encoder network for 2d medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 38, no. 10, pp. 2281–2292, 2019.

[14] D.-P. Fan, T. Zhou, G.-P. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Inf-net: Automatic COVID-19 lung infection segmentation from ct images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2626–2637, 2020.

[15] G.-P. Ji, Y.-C. Chou, D.-P. Fan, G. Chen, H. Fu, D. Jha, and L. Shao, "Progressively normalized self-attention network for video polyp segmentation," *arXiv preprint arXiv:2105.08468*, 2021.

[16] S. Li, X. Sui, X. Luo, X. Xu, L. Yong, and R. S. M. Goh, "Medical image segmentation using squeeze-and-expansion transformers," in *The 30th International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

[17] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *ICLR*, 2016.

[18] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[19] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[21] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[22] Y. Chen, D. Zhao, L. Lv, and Q. Zhang, "Multi-task learning for dangerous object detection in autonomous driving," *Information Sciences*, vol. 432, pp. 559–571, 2018.

[23] D. Cheng, R. Liao, S. Fidler, and R. Urtasun, "DARNet: Deep active ray network for building segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7431–7439.

[24] E. Xie, P. Sun, X. Song, W. Wang, X. Liu, D. Liang, C. Shen, and P. Luo, "Polarmask: Single shot instance segmentation with polar representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 193–12 202.

[25] U. Schmidt, M. Weigert, C. Broaddus, and G. Myers, "Cell detection with star-convex polygons," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 265–273.

[26] S. Wang, M. Liu, J. Lian, and D. Shen, "Boundary coding representation for organ segmentation in prostate cancer radiotherapy," *IEEE Transactions on Medical Imaging*, vol. 40, no. 1, pp. 310–320, 2020.

[27] Y. Meng, M. Wei, D. Gao, Y. Zhao, X. Yang, X. Huang, and Y. Zheng, "CNN-GCN aggregation enabled boundary regression for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 352–362.

[28] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1988.

[29] X. Chen, B. M. Williams, S. R. Vallabhaneni, G. Czanner, R. Williams, and Y. Zheng, "Learning active contour models for medical image segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 632–11 640.

[30] X. Chen, X. Luo, G. Wang, and Y. Zheng, "Deep elastica for image segmentation," in *Proceedings of the IEEE International Symposium on Biomedical Imaging*, 2021, pp. 706–710.

[31] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. B. Ayed, "Boundary loss for highly unbalanced segmentation," in *International Conference on Medical Imaging with Deep Learning*. PMLR, 2019, pp. 285–296.

[32] Z. Zhang, H. Fu, H. Dai, J. Shen, Y. Pang, and L. Shao, "ET-Net: A generic edge-attention guidance network for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 442–450.

[33] Y. Fang, C. Chen, Y. Yuan, and K.-Y. Tong, "Selective feature aggregation network with area-boundary constraints for polyp segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 302–310.

[34] S. Wang, L. Yu, K. Li, X. Yang, C.-W. Fu, and P.-A. Heng, "Boundary and entropy-driven adversarial learning for fundus image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 102–110.

[35] B. Murugesan, K. Sarveswaran, S. M. Shankaranarayana, K. Ram, J. Joseph, and M. Sivaprakasam, "Psi-Net: Shape and boundary aware joint multi-task deep network for medical image segmentation," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 7223–7226.

[36] B. Wang, W. Wei, S. Qiu, S. Wang, D. Li, and H. He, "Boundary aware U-Net for retinal layers segmentation in optical coherence tomography images," *IEEE Journal of Biomedical and Health Informatics*, 2021.

[37] R. Zhang, G. Li, Z. Li, S. Cui, D. Qian, and Y. Yu, "Adaptive context selection for polyp segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 253–262.

[38] X. Dong, J. Shen, L. Shao, and L. Van Gool, "Sub-Markov random walk for image segmentation," *IEEE Transactions on Image Processing*, vol. 25, no. 2, pp. 516–527, 2015.

[39] J. Shen, Y. Du, W. Wang, and X. Li, "Lazy random walks for superpixel segmentation," *IEEE Transactions on Image Processing*, vol. 23, no. 4, pp. 1451–1462, 2014.

[40] J. Shen, J. Peng, X. Dong, L. Shao, and F. Porikli, "Higher order energies for image segmentation," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4911–4922, 2017.

[41] J. Shen, X. Dong, J. Peng, X. Jin, L. Shao, and F. Porikli, "Submodular function optimization for motion clustering and image segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 9, pp. 2637–2649, 2019.

[42] J. Yao, J. Cai, D. Yang, D. Xu, and J. Huang, "Integrating 3D geometry of organ for improving medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 318–326.

[43] U. Wickramasinghe, E. Remelli, G. Knott, and P. Fua, "Voxel2mesh: 3D mesh model generation from volumetric data," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 299–308.

[44] S. Y. Shin, S. Lee, I. D. Yun, and K. M. Lee, "Deep vessel segmentation by learning graphical connectivity," *Medical Image Analysis*, vol. 58, p. 101556, 2019.

[45] Y. Meng, W. Meng, D. Gao, Y. Zhao, X. Yang, X. Huang, and Y. Zheng, "Regression of instance boundary by aggregated cnn and gcn," in *European Conference on Computer Vision*. Springer, 2020, pp. 190–207.

[46] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. H. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[47] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[48] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph CNN for learning on point clouds," *ACM Transactions on Graphics*, vol. 38, no. 5, pp. 1–12, 2019.

[49] A. Luo, X. Li, F. Yang, Z. Jiao, H. Cheng, and S. Lyu, "Cascade graph neural networks for rgb-d salient object detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 346–364.

[50] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, 2016, pp. 565–571.

[51] J. Ma, Z. Wei, Y. Zhang, Y. Wang, R. Lv, C. Zhu, C. Gaoxiang, J. Liu, C. Peng, L. Wang *et al.*, "How distance transform maps boost segmentation cnns: an empirical study," in *Medical Imaging with Deep Learning*. PMLR, 2020, pp. 479–492.

[52] D. A. Forsyth and J. Ponce, *Computer vision: a modern approach*. Pearson,, 2012.

[53] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "Kvasir-seg: A segmented polyp dataset,"

in *International Conference on Multimedia Modeling*. Springer, 2020, pp. 451–462.
[54] J. Sivaswamy, S. Krishnadas, G. D. Joshi, M. Jain, and A. U. S. Tabish, "Drishti-GS: Retinal image dataset for optic nerve head (ONH) segmentation," in *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2014, pp. 53–56.
[55] Z. Zhang, F. S. Yin, J. Liu, W. K. Wong, N. M. Tan, B. H. Lee, J. Cheng, and T. Y. Wong, "ORIGA-light: An online retinal fundus image database for glaucoma analysis and research," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, 2010, pp. 3065–3068.
[56] A. Almazroa, S. Alodhayb, E. Osman, E. Ramadan, M. Hummadi, M. Dlaim, M. Alkatee, K. Raahemifar, and V. Lakshminarayanan, "Retinal fundus images for glaucoma analysis: the RIGA dataset," in *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*, vol. 10579. International Society for Optics and Photonics, 2018, p. 105790B.
[57] F. Fumero, S. Alayón, J. L. Sanchez, J. Sigut, and M. Gonzalez-Hernandez, "RIM-ONE: An open retinal image database for optic nerve evaluation," in *24th International Symposium on Computer-based Medical Systems (CBMS)*. IEEE, 2011, pp. 1–6.
[58] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015.
[59] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE Transactions on Medical Imaging*, vol. 35, no. 2, pp. 630–644, 2015.
[60] D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, A. M. López, A. Romero, M. Drozdzal, and A. Courville, "A benchmark for endoluminal scene segmentation of colonoscopy images," *Journal of Healthcare Engineering*, vol. 2017, 2017.
[61] P. Kohli, P. H. Torr *et al.*, "Robust higher order potentials for enforcing label consistency," *International Journal of Computer Vision*, vol. 82, no. 3, pp. 302–324, 2009.
[62] G. Csurka, D. Larlus, F. Perronnin, and F. Meylan, "What is a good evaluation measure for semantic segmentation?." in *BMVC*, vol. 27, no. 2013, 2013, pp. 10–5244.
[63] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 724–732.
[64] Y. Chen, M. Rohrbach, Z. Yan, Y. Shuicheng, J. Feng, and Y. Kalantidis, "Graph-based global reasoning networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 433–442.
[65] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *ICLR*, 2017.
[66] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, Jun. 1947. [Online]. Available: https://doi.org/10.1007/bf02295996
[67] Y. Zhao, L. Rada, K. Chen, S. P. Harding, and Y. Zheng, "Automated vessel segmentation using infinite perimeter active contour model with hybrid region information with application to retinal images," *IEEE Transactions on Medical Imaging*, vol. 34, no. 9, pp. 1797–1807, 2015.
[68] Y. Ma, H. Hao, J. Xie, H. Fu, J. Zhang, J. Yang, Z. Wang, J. Liu, Y. Zheng, and Y. Zhao, "ROSE: A retinal OCT-angiography vessel segmentation dataset and new model," *IEEE Transactions on Medical Imaging*, vol. 40, no. 3, pp. 928–939, 2020.