Co-Net: A Collaborative Region-Contour-Driven Network for Fine-to-Finer Medical Image Segmentation

Anran Liu[1, 2 *], Xiangsheng Huang[2 †], Tong Li[1, 2 *], Pengcheng Ma[2, 3 †]
[1]School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China
[2]Institute of Automation, Chinese Academy of Science, Beijing, China
[3]Department of Radiation Oncology, Nanfang Hospital, Southern Medical University, Guangzhou, China

Abstract

In this paper, a ***fine-to-finer*** *segmentation task is investigated driven by region and contour features collaboratively on Glomerular Electron-Dense Deposits (GEDD) in view of the complementary nature of these two types of features. To this end, a novel network (Co-Net) is presented to dynamically use* ***fine*** *saliency segmentation to guide* ***finer*** *segmentation on boundaries. The whole architecture contains double mutually boosted decoders sharing one common encoder. Specifically, a new structure named Global-guided Interaction Module (GIM) is designed to effectively control the information flow and reduce redundancy in the cross-level feature fusion process. At the same time, the global features are used in it to make the features of each layer gain access to richer context, and a* ***fine*** *segmentation map is obtained initially; Discontinuous Boundary Supervision (DBS) strategy is applied to pay more attention to discontinuity positions and modifying segmentation errors on boundaries. At last, Selective Kernel (SK) is used for dynamical aggregation of the region and contour features to obtain a* ***finer*** *segmentation. Our proposed approach is evaluated on an independent GEDD dataset labeled by pathologists and also on open polyp datasets to test the generalization. Ablation studies show the effectiveness of different modules. On all datasets, our proposal achieves high segmentation accuracy and surpasses previous methods.*

1. Introduction

Glomerular disease is the main cause of kidney failure worldwide. It is now recognized that slow progression to end-stage renal disease occurs in up to 50% of affected patients [27]. Therefore, the importance of early diagnosis and treatment is accepted globally. Currently, the most effective examination is based on the identification of Glomerular Electron-Dense Deposits (GEDD) in scanning electron micrographs [14]. However, the complex representation of GEDD leads it time-consuming and into a dilemma for both pathologists and computers. As shown in Figure 1 (a), the GEDD often vary in space and appearance and have huge quantity differences and individual distinctions. Besides, with blurred boundary and fractured texture, they usually change over time [15]. Therefore, realizing accurate segmentation of GEDD is pressing and significant for auxiliary diagnosis.

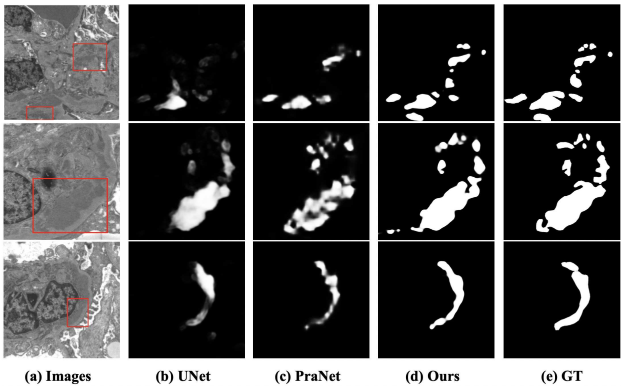

Figure 1. (a) Examples of GEDD diversity: GEDD drastically differ in size, shape and spacial representation and they have massive fracture and discontinuities; (b) and (c) are examples of GEDD segmentation produced by UNet and PraNet; (d) segmentation map generated by Co-Net. Previous models struggle at saliency information extraction and discontinuity boundary segmentation.

Recent years have witnessed the remarkable success of segmentation networks in various medical image applications. The Fully Convolutional Network (FCN) is a commonly used architecture to preserve the spatial information [25]. Inspired by FCN, UNet [30] achieved a more detailed

† Corresponding Authors. This work has been supported by the China National Key R&D Program (Grant No. 2020YFC2006406). *This work was done when Anran Liu and Tong Li were interns at Institute of Automation, Chinese Academy of Science, Beijing, China.

fusion of multi-scale features through skip concatenations on the encoder-decoder structure. Some following methods [40, 17, 36] aim to improve the architecture of U-net and achieve boosted performance. Despite some progress, these methods still face dilemmas when being confronted with tasks that require a high degree of precision, especially for complex spatial features and edge details. In fact, a finer segmentation is usually driven by both region and boundary information. Some works have been carried out based on each aspect. Some [3, 13] obtained regional saliency information by fusing cross-layer features, but the neglect of semantic gaps often affects the effectiveness of feature mining and fusion as pointed out in [37]. Then the gated mechanism [21] and global features [22] were considered to control information propagation and bridge semantic gap. In addition, several works introduced boundary information into networks by utilizing contour map [35, 11]. However, most of them still struggled at discontinuity positions and were not readily applicable to our task. Moreover, [28, 34] made use of both kinds of information; however, they exploited both the region and boundary information simultaneously in a single end-to-end framework without taking the differences between them into account. Therefore, it is still challenging to design a network driven by both region and boundary features and obtain a *fine-to-finer* result.

Unlike most existing methods, pathologists usually use as abundant region information as possible to guide contour modification to obtain finer segmentation results when facing the above challenges. Based on this, in this paper, a dynamic integrated network Co-Net is proposed with two boosted decoders for a *fine-to-finer* segmentation. Specifically, Co-Net mines the region features precisely and gets the segmentation through the *fine* decoder to guide the contour correction of the *finer* decoder. Meanwhile, an interactive mechanism is employed to promote both decoders to obtain a finer segmentation. As shown in Figure 2, the overall framework consists of three parts: (1) Global-guided Interaction Module (GIM) uses the dynamically extracted global features as guidance to control the progressive feature fusion in a gated manner, and initially obtains a fine segmentation map; (2) Discontinuous Boundary Supervision (DBS) with an attention-based module [5] further modifies segmentation map, which makes the model pay more attention to the boundary especially discontinuity positions; (3) Through using channel-wise attention [16], Selective Kernel (SK) integrates features from both decoders to achieve a finer segmentation.

The contributions of this paper are four-fold.

1. A novel collaborative region-contour-driven network is proposed to perform *fine-to-finer* segmentation of medical images.

2. We design a gated cross-level feature fusion module GIM to bridge the semantic gap to obtain the fine segmentation. A gated structure effectively controls the information flow to reduce redundancy, and the features of each layer gain access to richer context with the help of global features.

3. DBS with an attention-based module pays more attention to contour and modifies error on discontinuous textures under the guidance of fine segmentation above; SK achieves further integration.

4. Our model is evaluated on a new GEDD dataset and surpasses the aforementioned models by a large margin. The excellent performance on the polyp datasets also verifies the generalization of Co-Net.

2. Related Work

2.1. Medical Image Segmentation

Medical image segmentation task faces the challenges of mining finer details and lacking sufficient data compared with networks used in natural image segmentation such as SegNet [2] and DeepLab [4]. Since a U-shaped network called UNet [30] was proposed, it has been widely used in various medical image segmentation tasks like [26, 1] for its effective skip-connection. Further, many improvements have been made based on UNet. Among previous works, Zhou *et al*. [40] rebuilt nested networks with dense connections to capture aggregated features; ResUNet [36] introduced residual connection into UNet structure and in [7], Dolz *et al*. designed a multi-modal UNet to handle multi-scale contexts. Although the skip-connection can supplement information in the decoding process, a single decoder is not sufficient for segmentation tasks that need to consider the influence of multiple factors.

2.2. Region Context Features Fusion

In medical image segmentation tasks, cross-level feature fusion has received growing attention for effectively integrating region context information. Some fusion strategies progressively fused multi-scale features in a top-down manner, e.g. the feature pyramid network (FPN [23]), leading to information redundancy and dilution. Other works directly aggregate multi-scale features from different and distant layers. For example, in [22], Li *et al*. aggregated multi-scale features from different layers into feature maps that have access to both the high-and low-level information. However, in [37], Zhang *et al*. pointed out that it is less effective to directly integrate low-level and high-level semantic information, due to the semantic gap and spatial resolution. They proposed a new framework to bridge the gap between low-level and high-level features.

Based on their views, some novel fusion strategies were proposed. In recent work, Liu *et al*. [24] designed a relative

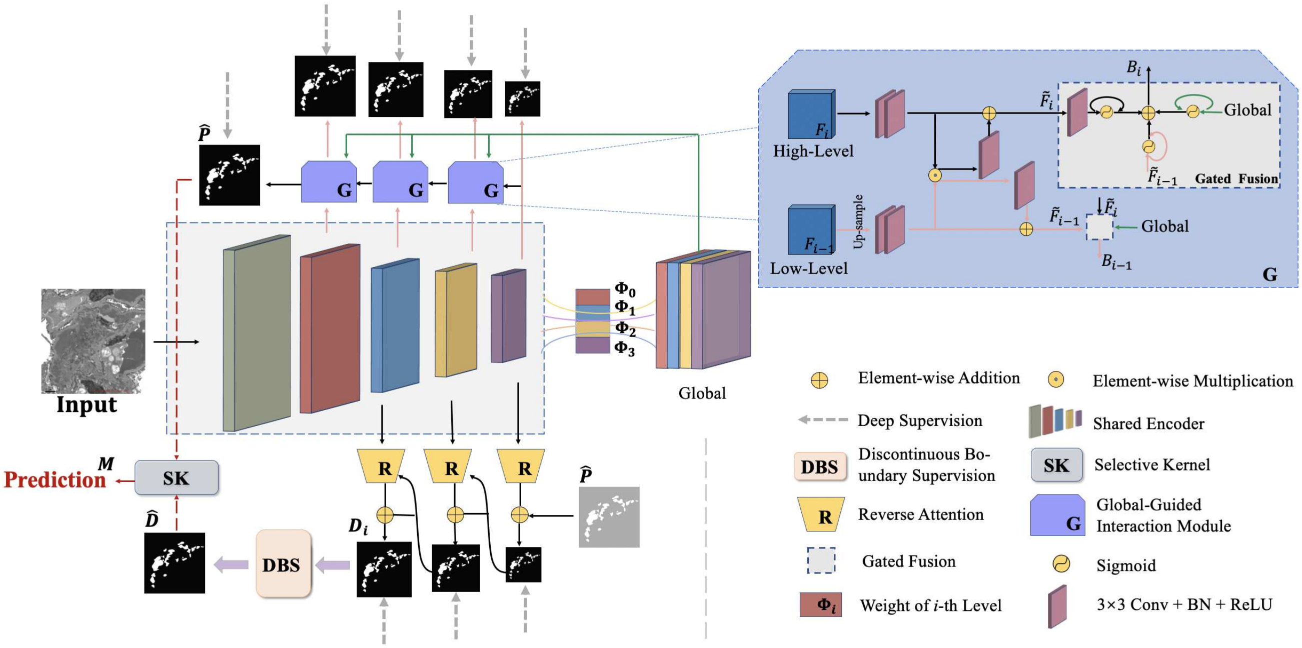

Figure 2. Overview of the proposed Co-Net with duplex decoders structure.

global calibration module to achieve the cross-scale information interaction. Li *et al.* [21] proposed a Gated Fully Fusion to selectively fuse features from multiple levels using gates. Inspired by these, in the process of cross-level feature fusion, our model uses a designed gated structure to control the information flow, and introduces global information as a guidance to bridge the semantic gap.

2.3. Boundary Refinement

Since the segmentation network is still facing a dilemma over accurate object edge segmentation, boundary information has drawn increasing attention. Zhao *et al.* [38] designed two modules to extract salient edge and salient object features respectively, and then fused these with a one-to-one guidance module. Zhou *et al.* [39] proposed an interactive transmission mechanism to guide the network to learn the correlation between the region and boundary features. Zhang *et al.* [35] utilized an edge guidance module in the early encoding layers to learn the edge-attention representations. Fang *et al.* [11] employed a light U-Net structure in an additional branch to extract the polyp's edge.

However, these works do not yet meet the requirements of the finer segmentation, especially for the segmentation of the electron density. Besides, though many works paid special attention to the correlation [38, 39] between the region and contour features, the segmentation errors caused by the fractured textures are hard to mitigate.

3. Our Approach

In this section, we propose a novel Collaborative Region-Contour-Driven Network (Co-Net) that can perform *fine-to-finer* segmentation. As shown in Figure 2, a global feature is dynamically extracted from sharing encoding layers with layer-wise weights. Taking the global feature as guidance, one new designed branch (Global-guided Interaction Module, GIM) gradually integrates cross-level features in a gated controlling way to initially obtain a fine segmentation. Further, Discontinuous Boundary Supervision (DBS) with an attention-based module modifies the segmentation map. Then, Selective Kernel (SK) integrates features to produce the finer output.

3.1. *Fine* Decoder with Global-Guided Interaction Module

The essential basic task in cross-level feature interaction is to aggregate useful information together. The progressive fusion of FPN [23] often causes dilution of semantic features. Further, ExFuse [37] was designed to bridge semantic gap between high- and low-level features. Considering that global features also have access to both semantics and fine details, in this paper, we design a novel gated structure to control cross-level feature interaction guided by global features to obtain a fine segmentation.

Specifically, features $\{F_i \in \mathbb{R}^{H_i \times W_i \times c_i}\}_{i=1}^{L}$ $(L = 5)$ are extracted from different layers of the encoder, where features are ordered by their depth in the network and H_i, W_i, c_i are the height, width, and channels' number

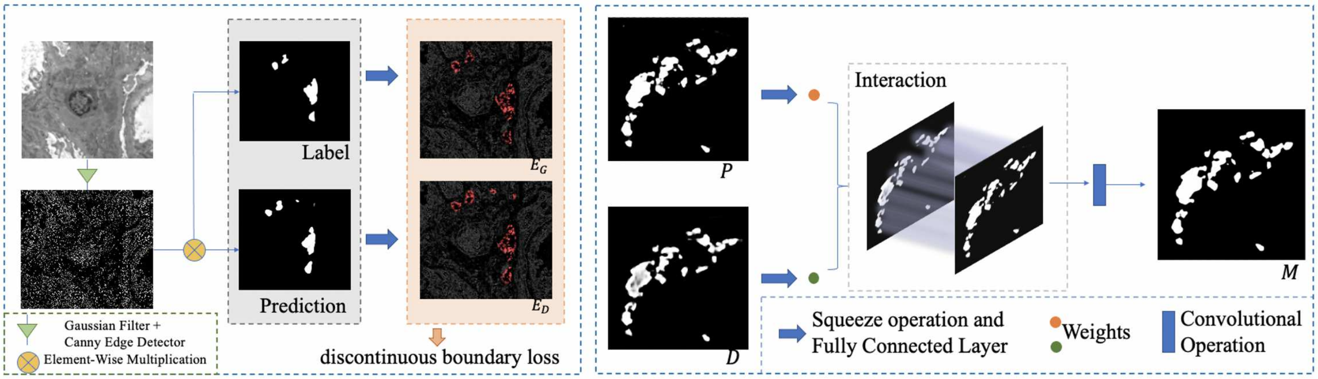
Label
Prediction
E_G
E_D
Gaussian Filter + Canny Edge Detector
Element-Wise Multiplication
discontinuous boundary loss
P
D
Interaction
M
Squeeze operation and Fully Connected Layer
Weights
Convolutional Operation

Figure 3. Core Modules: (left) Discontinuous Boundary Supervision (DBS); (right) Selective Kernel (SK).

of F_i, respectively. Inspired by SENet [16], we align the layer level of the encoding feature with the channel level in the SE block, and implement the Squeeze-and-Excitation operation to each layer to learn the correlation between different layers, giving different attention representations to them. We apply global average pooling to encoding features $\{F_i\}, i \in \{2, 3, 4, 5\}$, respectively, and further concatenate them to construct a fused multi-scale representation $\mathbf{Q} \in \mathbb{R}^{C\times 1}$, $C = \sum_{i=2}^{5} c_i$. A layer-wise fusion weight $\Phi \in \mathbb{R}^{1\times 4}$ is obtained through operating two fully connected layers over $\mathbf{Q}$. Further, the original features can be updated by $\hat{F}_i = F_i * \Phi_i$. With the help of series upsampling and concatenation operation, the global context feature can be generated as

$$\mathcal{F} = \hat{F}_2 \oplus UP(\hat{F}_3) \oplus UP(\hat{F}_4) \oplus UP(\hat{F}_5). \tag{1}$$

In order to give full play to the guidance and supplementary role of global feature $\mathcal{F}$, and to make the progressive integration more refined, we design the gated progressive feature fusion, in which the feature of each layer has access to $\mathcal{F}$ and the multi-level features can be integrated steadily (the effectiveness is shown in 4.3).

Specifically, as shown in Figure 2, a fusion mechanism is firstly applied between adjacent levels. The combinations of 3 × 3 convolution, BatchNorm and ReLU (V) are associated with features (F_{i-1}, F_i) respectively to fit them in and obtain the consistent further features. After that, the multiplication-fusion is formally defined as $U = V_{i-1}(F_{i-1}) \cdot V_i(F_i)$; it helps the fusion outcome combine the properties of both the F_{i-1} and F_i and suppress redundancy on both sides. The addition-refining followed, $\tilde{F} = V(F + V(U))$, in which some useful information can be regulated again. Then, a gated structure is designed to further integrate features under the interference of $\mathcal{F}$. We perform sigmoid operations over $\tilde{F}_i$, $\tilde{F}_{i-1}$ and global feature $\mathcal{F}$ to get their respective gate maps A_i, A_{i-1} and A_g. Then, addition-based fusions with different weight configurations are associated with $\tilde{F}_i$ and $\tilde{F}_{i-1}$ respectively, which are

$$\begin{aligned} B_i &= (1 + A_i) \cdot \tilde{F}_i + (1 - A_i)(A_{i-1} \cdot \tilde{F}_{i-1} + A_g \cdot \mathcal{F}) \\ B_{i-1} &= (1 + A_{i-1}) \cdot \tilde{F}_{i-1} + (1 - A_{i-1})(A_i \cdot \tilde{F}_i + A_g \cdot \mathcal{F}) \end{aligned} \tag{2}$$

where $+$ means element-wise addition and $\cdot$ denotes element-wise multiplication broadcasting in the channel dimension. As shown in Figure 2, B_{i-1} is used for deep supervision and B_i is the input of the next GIM.

3.2. *Finer* Decoder with Discontinuous Boundary Supervision

It is worth noting that in the clinical diagnosis of complex pathological images, the pathologist locates the suspicious area and then inspects local tissue to accurately distinguish the deposits' contour and pattern. Inspired by this, we design a *finer* decoder that is complementary to the *fine* decoder and focuses more on the contour. With the help of fine segmentation $\hat{P}$, the sequential attention modules guide the network to give more attention to the segmentation boundary to make up for the deficiency of the *fine* decoder. Then DBS is used to modify discontinuous positions in pixel level.

To achieve this, same as [5], we utilize the Reverse Attention (RA) mechanism to erase the current predicted salient regions in the features and guide the network to exploit the missing details. To further amend the boundaries, we propose a fine-grained approach DBS to mitigate the fracture misalignment. Specifically, applying some traditional edge detection methods (Gaussian filter, Canny edge detector) to the row images, we firstly get the edge maps. As shown in Figure 3 (left), through multiplying the edge maps by our segmentation maps and masks, respectively, E_D and E_G are obtained. Then the DBS module can leverage additional pixel-level supervision to modify errors on discontinuity .

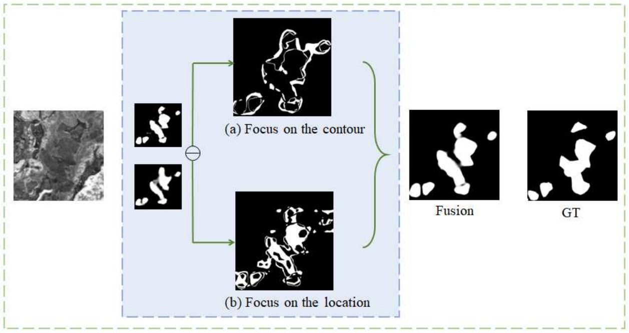

Figure 4. The attention area of region locating feature and contour correction feature.

3.3. Correlation Analysis & Selective Kernel

In order to increase the interpretability of our model, we first visualize the attention area of region locating feature from *fine* decoder and contour correction feature from *finer* decoder to analyze the correlation of them.

Figure 4 (a) shows the result of subtracting $\hat{P}$ from $\hat{D}$ and (b) shows the opposite result. Compared (a) and (b), we can see that the $\hat{D}$, going through the DBS, precisely highlights the segmentation around the contour and discontinuity positions more. While $\hat{P}$ pays more attention to salient object regions. Thus, during the *fine-to-finer* segmentation process of GEDD, acting like a human vision system, the model pays different levels of attention to object region and boundary, which are actually not independent.

To finally achieve a dynamic integration of features ($\hat{P}$ and $\hat{D}$) obtained from two decoders, we employ a simplified SE block from the SENet [16].

In fact, the cross-level region features $\{P_j\}_{j=2}^{5}$ from *fine* decoder and the discontinuous boundary features from *finer* decoder are not independent. Acting like human experts, our model pays more attention to cross-level region features from the start to locate the dense stuff accurately. Then it needs more discontinuous boundary information to correct the wrong depiction of edges with the help of DBS. As shown in Figure 3 (right), the input features $\hat{D}$ and $\hat{P}$ are firstly concatenated as $S \in \mathbb{R}^{W \times H \times C}$, h_c is the feature among each channel, and SK shrinks S in spatial dimensions as:

$$z_c = \frac{1}{W \times H} \sum_{i=1}^{H} \sum_{j=1}^{W} h_c(i,j), \tag{3}$$

where $z = \{z_1, \cdots, z_C\} \in \mathbb{R}^C$, $S = \{s_1, \cdots, s_C\}$. Then the channel-wise weights are calculated through performing fully connected layers over z. Multiplying the weights to the corresponding features, we get a *finer* prediction M.

3.4. Multi-Level Joint Loss Function

In order to optimize region structure and boundary detail segmentation collectively, our loss function is composed of two parts: discontinuous boundary loss L_{disc}, and multi-scale integrated feature loss L_{inte}.

In [29] Qin *et al*. designed the weighted IoU loss and binary cross entropy (BCE) loss for highlighting the importance of some hard pixels as $\mathcal{L} = \mathcal{L}_{IoU}^{w} + \mathcal{L}_{BCE}^{w}$. However, restricted on image element, the pixel-level loss neglects the crucial structural information. [33] proposed a new $ms\text{-}ssim$ loss; it can not only concentrate on the similarity of the structure, but also provide different weights for different resolutions. Therefore, our basic loss is denoted as $\mathcal{L} = \mathcal{L}_{IoU}^{w} + \mathcal{L}_{BCE}^{w} + \mathcal{L}_{MS-SSIM}$. Considering the DBS strategy, we design $\mathcal{L}_{disc}$ as:

$$\begin{aligned} \mathcal{L}_{disc} = & \lambda_1 \mathcal{L}(G, \hat{P}_{up}) + \lambda_2 \mathcal{L}(G, M) + \lambda_3 \sum_{i=3}^{5} \mathcal{L}(G, D_i^{up}) \\ & + \lambda_4 \mathcal{L}_{BCE}^{w}(E_G, E_D), \end{aligned} \tag{4}$$

where $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 0.25$, $\lambda_4 = 0.15$ and G denotes ground truth, $\hat{P}$, M, D_i, E_g and E_d are features noted in Figure 2 and Figure 3.

As shown in Figure 2, the supervision of multi-scale features generated by the GIM can be calculated as $\mathcal{L}_{inte} = \sum_{j=2}^{5} \frac{1}{2^{j-1}} \mathcal{L}(G, P_j^{up})$, where higher level loss has smaller weight due to the larger error existing in that higher features. Thus the final loss function is defined as: $\mathcal{L}_{total} = \mathcal{L}_{disc} + \mathcal{L}_{inte}$.

4. Experiments

To demonstrate the effectiveness and generalization of the proposed method, we conduct experiments on GEDD dataset and several open datasets of polyp segmentation.

4.1. Dataset

We note that there is no public GEDD dataset, a serious drawback to the research on diagnosing Glomerular diseases. Accordingly, in this paper, some electron micrographs are collected from the clinical diagnosis, which contains IgA nephropathy (IgA), Lupus Nephritis (LN) and Membranous Nephropathy(MN). They are all grayscale images of 2048 × 2048. We keep 217 images in the dataset after the pathologists' strict examination. 80% of the dataset is randomly selected as the training set and the remaining 20% as the test set. Table 1 shows the statistics.

Type	Cases (%)	Gender (M:F)	Age	Images/Case
IgA	46.0%	5:4	7∼69	1∼3
LN	17.4%	1:6	10∼67	1∼4
MN	36.6%	8:5	24∼75	1∼4

Table 1. Statistics of GEDD dataset.

Besides, for generalization evaluation, we conduct extra experiments on five polyp segmentation datasets, ETIS,

Methods	mean Dice	mean IoU	F_β^w	S_α	E_ϕ^{max}	MAE	Infer time
UNet	0.605	0.462	0.531	0.775	0.922	0.018	22.43
UNet++	0.385	0.262	0.333	0.596	0.829	0.029	28.10
SegNet	0.162	0.091	0.161	0.516	0.528	0.026	21.27
ResUNet	0.439	0.294	0.363	0.649	0.847	0.026	43.61
PraNet	0.584	0.430	0.535	0.759	0.951	0.018	40.38
Co-Net(Ours)	**0.797**	**0.669**	**0.761**	**0.851**	**0.978**	**0.010**	42.45
Δ	+31.7%	+44.8%	+42.2%	+9.8%	+2.8%	+44.4%	-

Table 2. Comparison with different state-of-the-art methods on GEDD dataset. The basic evaluation metrics, mean Dice, mean IoU and MAE, are significantly higher for Co-Net compared to other competing models. Meanwhile, our model is also outstanding in F_β^w, S_α and E_ϕ^{max} that can further evaluate the pixel-level and global-level similarity. The Infer time (ms) is calculated by averaging over whole test dataset with mini batch size 1 and an image of 2048 × 2048 resolution. We see that Co-Net has an advantage over other models as balancing accuracy and time consume well. The accuracy percentage increase of our model compared to the state-of-the-art results is represented by Δ. (The best result is highlighted in bold and the second best result is underlined.)

Methods	CVC-300			CVC-ColonDB			ETIS-LaribPolypDB		
	mean Dice	mean IoU	Infer.	mean Dice	mean IoU	Infer.	mean Dice	mean IoU	Infer.
UNet	0.677	0.561	23.88	0.510	0.441	20.92	0.398	0.304	22.97
UNet++	0.610	0.503	31.00	0.494	0.381	26.57	0.303	0.227	29.93
SegNet	0.778	0.671	19.86	0.501	0.420	19.36	0.285	0.226	21.33
ResUNet	0.298	0.200	43.71	0.342	0.248	41.89	0.266	0.184	44.03
PraNet	0.871	0.797	38.39	**0.709**	**0.640**	35.71	0.628	0.567	37.15
Co-Net(Ours)	**0.872**	**0.798**	39.61	0.687	0.618	37.98	**0.641**	**0.582**	39.54

Table 3. Comparison with different state-of-the-art methods on three polyp datasets (CVC-300, CVC-ColonDB and ETIS-LaribPolypDB).

CVC-ClinicDB, CVC-ColonDB, EndoScene, and Kvasir [19, 18, 31, 32, 20]. Then we follow the same training settings in PraNet [10]. The images from Kvasir, and CVC-ClinicDB are used for training and the total number is 1450. We take PraNet as a reference and construct extra testing data composed of images from CVC-ColoDB (380 images), ETIS (196 images) and EndoScecne (60 images).

4.2. Implementation Details and Comparative Experiments

Implementation Details. Due to the GEDD images' high-resolution, we uniformly set all images to a fixed size of 512 × 512. The initial weights of the encoder network come from Res2Net [12] pretrained on ImageNet. A multi-scale training strategy $\{0.75,\ 1,\ 1.25\}$ is used in training. We set the $batch\_size$ to 4 with synchronized batch normalization and adopt Adam optimizer with a $learning\_rate$ of $1e{-}4$. Our Co-Net is based on Pytorch framework. An end-to-end training process (50 $epochs$) is trained with NVIDIA TITAN Xp GPU. During inferring, we resize each image to 512 × 512 and feed it to Co-Net with batch size 1 to predict.

In the polyp segmentation task, the implementation details are the same as GEDD segmentation, except that the training and inferring size is set to 352, the training batch size is 16 and the epoch number is 100.

Evaluation Metrics. We use mean Dice, mean IoU and MAE (mean absolute error) for comprehensively quantitative evaluation [6]. Besides, as in [8, 9], giving further consideration to the model performance, we employ "F_β^w" (weighted F_β measure), "S_α" (structural similarity metric) and "E_ϕ^{max}" (enhanced-alignment metric).

Comparative Experiments. We compare our Co-Net with UNet [30], UNet++ [40], SegNet [2], ResUNet [36] and PraNet [10]. As shown in Table 2, our method outperforms five state-of-the-art methods on all metrics on GEDD. And some visualization results are shown in Figure 5. SegNet does not exhibit a competitive result. The reason is supposed because of the lack of substantial details without skip-connection. And the U-shaped network bridges this gap and concatenations are effective. However, the comparison of UNet and UNet++ indicates overly dense concatenations lead to performance degradation. Furthermore, although UNet roughly achieves parity with PraNet on quantitative results, the visualized figure shows that PraNet performs better in detailed boundary processing, which also indicates the effectiveness of the boundary information. Moreover, ResUNet is incapable of extracting abundant contextual information in its encoder, resulting in a significante accuracy drop. Accordingly, we employ Res2Net as sharing encoder to enhance the feature robustness.

Table 3 shows the segmentation results on polyp datasets and some detailed results visualization are provided in Figure 6. Our model still has a stable performance on the polyp datasets. Especially in many challenging cases, such as in

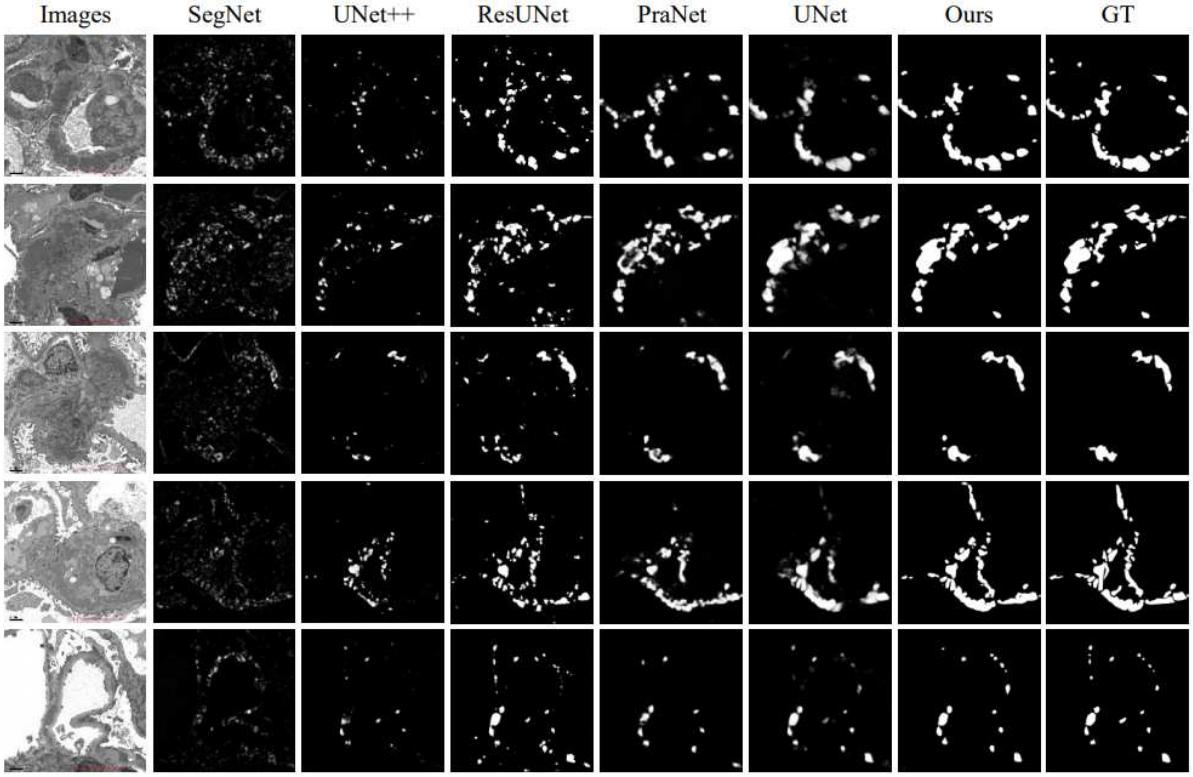

Figure 5. GEDD segmentation results of different methods.

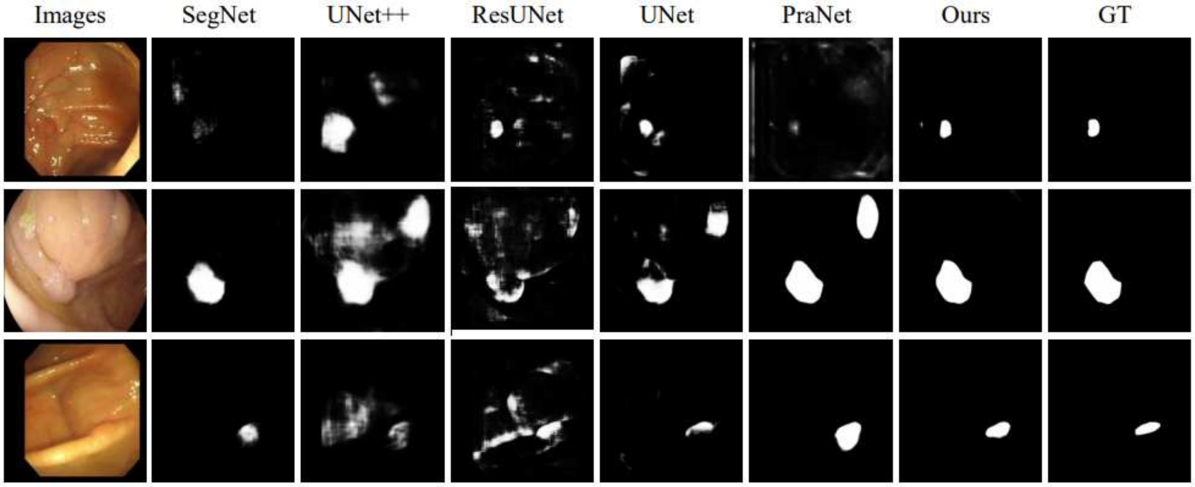

Figure 6. Polyp segmentation results of different methods.

ETIS dataset, our model has advantages in the capture of complex spatial features and the refinement of complex textures. We can also see in Table 2 and 3 that our model has provided a good balance between accuracy and speed.

4.3. Ablation Study

To evaluate the effectiveness and interactivity of each module of the proposed method, we compare Co-Net with its four variants on GEDD dataset in Table 1 to provide deeper insight into our model.

Specifically, the baseline model refers to a U-shaped framework with RA module [10]. We add different components to the baseline in turn to investigate their effectiveness. As shown in Table 4, “Baseline+GIM” only adds GIMs on the baseline; “Baseline+DBS” only applies DBS to the baseline; “Baseline+GIM+DBS” adds GIM and DBS

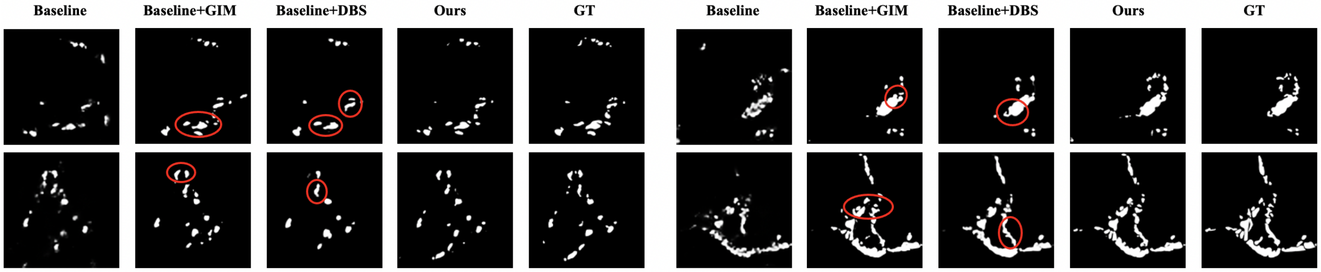

Figure 7. GEDD segmentation results of ablation study (Failed segmentations are marked in red circles).

Methods	mean Dice	mean IoU	F_β^w	S_α	E_ϕ^{max}	MAE
Baseline	0.584	0.430	0.535	0.759	0.951	0.018
+GIM	0.763(↑ 30.7%)	0.625(↑ 45.3%)	0.715(↑ 33.6%)	0.832(↑ 9.6%)	0.968(↑ 1.8%)	0.011(↑ 38.9%)
+DBS	0.747(↑ 27.9%)	0.604(↑ 40.5%)	0.703(↑ 31.4%)	0.841(↑ 10.8%)	0.973(↑ 2.3%)	0.012(↑ 33.3%)
+GIM+DBS	0.731(↑ 25.2%)	0.588(↑ 36.7%)	0.684(↑ 27.9%)	0.830(↑ 9.4%)	0.951(-)	0.012(↑ 33.3%)
Ours	**0.797(↑ 36.5%)**	**0.669(↑ 55.6%)**	**0.761(↑ 42.2%)**	**0.851(↑ 12.1%)**	**0.978(↑ 2.8%)**	**0.010(↑ 44.4%)**

Table 4. Ablation study on GEDD dataset. Both "+GIM" and "+DBS" have high segmentation accuracy indicating the need to use region and contour information. The simple combination of two modules ("+GIM+DBS") brings worse results than the use of any single module, proving the influence of information redundancy. Our model further shows the effectiveness of SK.

modules to the baseline at the same time, and simply concatenates the two outputs as a prediction. By comparing them with "Baseline", it is apparent that each separate module can significantly improve segmentation accuracy, especially mean Dice and mean IoU show an increment of around 30%. Further, as shown in Figure 7, according to "Baseline+DBS", "Baseline" and Ours, we can see the DBS strategy is necessary for modifying the error caused by the discontinuous texture. It is helpful to guide model to achieve a finer segmentation. The different results of "Baseline+GIM", "Baseline" and Ours suggest that this module enables the network to more accurately capture structural information and finely segment salient region. By comparing "Baseline+GIM+DBS", "Baseline" and Ours, it can be seen that directly concatenating the outputs will cause information redundancy and performance degradation, and the SK module can effectively alleviate this problem.

Module Variants	GEDD		
	mean Dice	mean IoU	MAE
Ours	0.797	0.669	0.01
(A)	0.763	0.625	0.012
(B)	0.769	0.627	0.011
(C)	0.770	0.633	0.011

Table 5. The effectiveness of gated mechanism and global feature.

To assess the effectiveness of the combination of each component in GIM, we conduct extra experiments. First, we try to temporarily remove the progressive fusion and gate structure in GIM, and transmit the global feature $\mathcal{F}$ to RA (A). Then, we simply concatenate global feature and two adjacent features $\tilde{F}_{i-1}$ and $\tilde{F}_i$ to replace gated fusion (B). Also, the synthesis mechanism of the global feature has been changed, such as direct concatenation (C). Table 5 provides result comparisons. It is apparent that neither relying on global features alone nor directly integrating encoding features as global guidance is not satisfactory. Besides, the meticulous gated structure fusion is indeed helpful for the global feature to play a better guiding role and it is conducive to better fusion.

5. Conclusion

In this paper, we proposed a novel collaborative region-contour-driven Network (Co-Net) for the fine segmentation task. This network, driven by region information and boundary information, performed *fine-to-finer* segmentation. Taking the correlation of the region and contour information into account, the network extracted multi-scale features through a sharing encoder, and first generated a global feature to guide a progressive cross-level feature fusion in GIMs with an information filtering gate of *fine* decoder. Meanwhile, the *fine* segmentation obtained above was transmitted towards the *finer* decoder. Through DBS with an attention based module, the inaccurate segmentation at discontinuity positions drew more attention and was revised in pixel-level. Benefiting from these two collaborative decoders, the *fine-to-finer* segmentation pattern driven by region-contour feature, simulating people vision system, was a dynamic and integrated approach. Extensive experiments on both the independent GEDD dataset and open polyp segmentation datasets well demonstrated the effectiveness and robustness of the proposed network.

References

[1] Simon A.A.Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R. Ledsam, Klaus H. Maier-Hein, S.M.Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. *arXiv preprint arXiv: 1806.05034v4*, 2018.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *TPAMI*, 39(12):2481–2495, 2017.

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *Computer Science*, (4):357–361, 2014.

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 804–818, 2018.

[5] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. Reverse attention for salient object detection. In *ECCV*, 2018.

[6] Jha D., Smedsrud P.H., Riegler M.A., Johansen D., De Lange T., Halvorsen P., and Johansen H.D. Resunet++: An advanced architecture for medical image segmentation. *IEEE ISM*, 2019.

[7] Jose Dolz, Christian Desrosiers, and Ismail Ben Ayed. Ivdnet: Intervertebral disc localization and segmentation in mri with a multi-modal u-net. *Computational Methods and Clinical Applications for Spine Imaging*, pages 130–143, 2019.

[8] Fan D.P., Gong C., Cao Y., Ren B., Cheng M.M., and Borji A. Enhanced alignment measure for binary foreground map evaluation. In *IJCAI*, pages 442–450, 2018.

[9] Fan D.P., Cheng M.M., Liu Y., Li T., and Borji A. Structure-measure: A new way to evaluate foreground maps. In *IEEE ICCV*, pages 4548–4557, 2017.

[10] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *MICCAI*, pages 263–273, 2019.

[11] Yuqi Fang, Cheng Chen, Yixuan Yuan, and Kai yu Tong. Selective feature aggregation network with area-boundary constraints for polyp segmentation. In *MICCAI*, pages 302–310, 2019.

[12] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligenge*, pages 652–662, 2009.

[13] Golnaz Ghiasi and Charless C. Fowlkes. Laplacian pyramid reconstruction and refinement for semantic segmentation. In *ECCV*, pages 519–534, 2016.

[14] Mark Haas, Surya V. Seshan, Laura Barisoni, Kerstin Amann, Ingeborg M. Bajema, Jan Ulrich Becker, Kensuke Joh, Danica Ljubanovic, Ian S.D. Roberts, Joris J. Roelofs, Sanjeev Sethi, Caihong Zeng, and J. Charles Jennette. Consensus definitions for glomerular lesions by light and electron microscopy: recommendations from a working group of the renal pathology society. In *Kidney International*, pages 1120–1134, 2018.

[15] Renée Habib, Eric Girardin, Marie-France Gagnadoux, Nicole Hinglais, Micheline Levy, and Michel Broyer. Immunopathological findings in idiopathic nephrosis: Clinical significance of glomerular “immune deposits”. In *Pediatric Nephrology*, pages 402–408, 1988.

[16] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze and excitation networks. In *CVPR*, pages 7132–7141, 2018.

[17] Huimin Huang, Lanfen Lin1, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full–scale connected unet for medical image segmentation. In *ICASSP*, 2020.

[18] Bernal J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilarinaño, and F. Wmdova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *CMIG*, 43:99–111, 2015.

[19] Silva J., Histace A., Romain O., Dray X., and Granado B. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International Journal of Computer Assisted Radiology and Surgery*, 9(2):283–293, 2014.

[20] D. Jha, Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., Johansen, and H.D. Kvasir-seg: A segmented polyp dataset. In *MMM*, pages 451–462, 2020.

[21] Xiangtai Li, Houlong Zhao, Lei Han, Yunhai Tong, Shaohua Tan, and Kuiyuan Yang. Gated fully fusion for semantic segmentation. In *AAAI*, page 11418.

[22] Zun Li, Congyan Lang, Jun Hao Liew, Yidong Li, and Qibin Houand Jiashi Feng. Cross-layer feature pyramid network for salient object detection. *IEEE Transactions on Image Processing*, 30:4587–4598, 2021.

[23] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017.

[24] Jiang-Jiang Liu, Zhi-Ang Liu, and Ming-Ming Cheng. Centralized information interaction for salient object detection. *arXiv preprint arXiv: 2012.11294v2*, 2020.

[25] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 640–651, 2015.

[26] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *MICCAL*, pages 424–432, 2016.

[27] M. O’Shaughnessy, S. Hogan, Bawana D Thompson, R. Coppo, A. Fogo, and J. Jennette. Glomerular disease frequencies by race, sex and region: results from the international kidney biopsy survey. In *Nephrology Dialysis Transplantation*, pages 661–669, 2018.

[28] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *CVPR*, pages 9413–9422, 2020.

[29] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: boundary–aware salient object detection. In *CVPR*, pages 7479–7489, 2019.

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAL*, pages 234–241, 2015.

[31] Tajbakhsh, N., Gurudu, S.R., Liang, and J. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE TMI*, 35(2):630C644, 2015.

[32] Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., López, A.M., Romero, A., Drozdzal, M., Courville, and A. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of Healthcare Engineering*, 2017.

[33] Zhou Wang, Eero P. Simoncelli1, and Alan C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, pages 7132–7141, 2003.

[34] Jun Wei, Shuhui Wang, Zhe Wu, Chi Su, Qingming Huang, and Qi Tian. Label decoupling framework for salient object detection. In *CVPR*, pages 13025–13034, 2020.

[35] Zhijie Zhang, Huazhu Fu, Hang Dai, Jianbing Shen, Yanwei Pang, and Ling Shao. Et-net: A generic edge-attention guidance network for medical image segmentation. In *MICCAI*, pages 442–450, 2019.

[36] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018.

[37] Zhenli Zhang, Xiangyu Zhang, Chao Peng, Xiangyang Xue, and Jian Sun. Exfuse: Enhancing feature fusion for semantic segmentation. In *ECCV*, pages 269–284, 2018.

[38] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Ju-Feng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. In *ICCV*, pages 8779–8788, 2019.

[39] Huajun Zhou, Xiaohua Xie, Jianhuang Lai, Zixuan Chen, and Lingxiao Yang. Interactive two-stream decoder for accurate and fast saliency detection. In *CVPR*, pages 9141–9150, 2020.

[40] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested ucnet architecture for medical image segmentation. *IEEE TMI*, pages 3–11, 2019.