Precise Yet Efficient Semantic Calibration and Refinement in ConvNets for Real-time Polyp Segmentation from Colonoscopy Videos

Huisi Wu,[1*] **Jiafu Zhong,**[1] **Wei Wang,**[1] **Zhenkun Wen,**[1] **Jing Qin**[2]
[1] College of Computer Science and Software Engineering, Shenzhen University
[2] Centre for Smart Health, The Hong Kong Polytechnic University
hswu@szu.edu.cn

Abstract

We propose a novel convolutional neural network (ConvNet) equipped with two new semantic calibration and refinement approaches for automatic polyp segmentation from colonoscopy videos. While ConvNets set state-of-the-are performance for this task, it is still difficult to achieve satisfactory results in a real-time manner, which is a necessity in clinical practice. The main obstacle is the huge semantic gap between high-level features and low-level features, making it difficult to take full advantage of complementary semantic information contained in these hierarchical features. Compared with existing solutions, which either directly aggregate these features without considering the semantic gap or employ sophisticated non-local modeling techniques to refine semantic information by introduce many extra computational costs, the proposed ConvNet is able to more precisely yet efficiently calibrate and refine semantic information for better segmentation performance without increasing model complexity; we call the proposed ConvNet as *SCR-Net*, which has two key modules. We first propose a semantic calibration module (SCM) to effectively transmit the semantic information from high-level layers to low-level layers by learning the semantic-spatial relations during the training procedure. We then propose a semantic refinement module (SRM) to, based on the features calibrated by SCM, enhance the discrimination capability of the features for targeting objects. Extensive experiments on the Kvasir-SEG dataset demonstrate that the proposed SCR-Net is capable of achieving better segmentation accuracy than state-of-the-art approaches with a faster speed. The proposed techniques are general enough to be applied to similar applications where precise and efficient multi-level feature fusion is critical. The code is available at https://github.com/jiafuz/SCR-Net.

Introduction

Colorectal cancer (CRC) is a common malignant tumor in the gastrointestinal tract. Fortunately, we can effectively prevent the CRC if the colon polyps, the masses bulging on the surface of colon, are removed in time before developing to the CRC (Kolligs 2016). Colonoscopy is the primary method for prevention of colon cancer, in which a tiny camera is navigated into the colon in order to locate and remove polyps.

*Corresponding author. Email: hswu@szu.edu.cn.

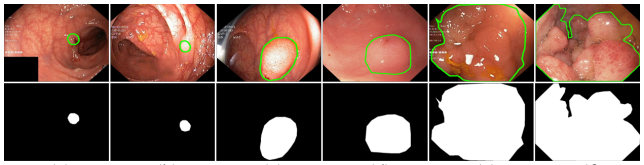

Figure 1: Illustration of challenges in automatic polyp segmentation: (a)-(b) small polyps, (c)-(d) middle polyps, and (e)-(f) large polyps, while (b), (d) and (e) show the polyps with low contrast to the background.

However, according to the study reported in (Leufkens et al. 2012), one out of every four polyps was missed during the colonoscopy procedures. Therefore, automatic approach to identifying and segmenting polyps from colonoscopy videos is highly demanded to improve the outcome of colonoscopy.

Precisely segment polyps from colonoscopy videos is indeed a challenging task. First, the variation of polyps is quite large; see Fig. 1 (a)-(f). Second, the low intensity contrast between polyp and background may worsen the discrimination of foreground features and hence increase the probability of incorrect segmentation, see Fig. 1 (b) and (d)-(e). Third, the segmentation should be carried out in a real-time manner, as it is anticipated that the results can be presented to doctors immediately for their prompt action during the colonoscopy procedures.

Early works for this task employ hand-crafted features to distinguish polyps from background, including intensity distribution, geometric features, and volumetric properties (Jerebko, Franaszek, and Summers 2002; Jerebko et al. 2003; Yao et al. 2004; Gross et al. 2009). However, the representative capability of these hand-crafted features is far from sufficient to meet above-mentioned challenges. In recent years, with the development of deep learning, more and more effort has been dedicated to using convolutional neural networks (CNNs) to handle this challenging task. For example, Li et al. (Li et al. 2017) and Brandao et al. (Brandao et al. 2017) propose to use fully convolutional networks (FCNs) to segment the polyps. However, the detailed boundary information is always missing in FCN architecture due to a series of down-sampling operations. Sun et al. (Sun et al. 2019) propose to improve the feature representation ability by in-

troducing the dilated convolution(Yu and Koltun 2015), but it is difficult to find suitable receptive fields to capture the appropriate contexts. ResU-Net++(Jha et al. 2019) attempts to capture multi-scale contexts by introducing atrous spatial pyramid pooling (Chen et al. 2017), but the multi-scale contexts still cannot be fully harnessed due to the existing of semantic gap between high-level and low-level features. Pra-Net (Fan et al. 2020) proposes a reverse attention module to calibrate misaligned predictions, but it still cannot effectively model the global context to deal with large variation of polyps. Overall, existing models are incapable of meeting above-mentioned challenges and, simultaneously, maintaining real-time performance to fulfill the clinical requirement.

In this paper, we propose a novel ConvNet equipped with two new semantic calibration and refinement approaches to meet these challenges. Compared with existing solutions, which either directly aggregate low-level and high-level features without considering the semantic gap or employ sophisticated non-local modeling techniques to refine semantic information by introducing many extra computational costs, the proposed ConvNet, is able to more precisely yet efficiently calibrate and refine semantic information for better segmentation performance without increasing model complexity; we call the proposed ConvNet as *SCR-Net*. The proposed *SCR-Net* has two key components. We first propose a semantic calibration module (SCM) to effectively transmit the semantic information from high-level layers to low-level layers by learning the semantic-spatial relations during the training procedure, which is able to greatly alleviate the semantic shift problem that cause blurring and implausible boundaries. We further propose a semantic refinement module (SRM) to, based on the features calibrated by SCM, enhance the discrimination capability of the features for targeting objects. Extensive experiments on the Kvasir-SEG dataset (Jha et al. 2020) demonstrate the effectiveness of the proposed method.

Our contributions can be summarized as:

- We propose a novel ConvNet to accurately segment polyps from colonoscopy videos in a real-time manner, which has potential to be applied to colonoscopy examination and improve the outcomes of this intervention.
- We propose two novel and efficient semantic calibration and refinement approaches to bridge the semantic gap between feature maps with different levels and hence take full advantage of the complementariness of these features to boost the segmentation performance; these two approaches are general enough to be harnessed in applications with similar challenges.
- We evaluate the proposed ConvNet on a popular dataset and experimental results demonstrate the proposed ConvNet achieves better segmentation results than state-of-the-art approaches with a faster speed.

Related Work

Polyp Segmentation

Early studies for this task are mainly based on various hand-crafted features. Nappi et al. (Nappi and Yoshida 2002) and Yoshida et al. (Yoshida et al. 2002) first relied on hysteresis thresholding with volumetric properties (such as shape and curvedness) to identify polyps. Jianhua et al. (Yao et al. 2004) also employed fuzzy clustering and deformable models for polyp segmentation. Sebastian et al. (Gross et al. 2009) and Ganz et al. (Ganz, Yang, and Slabaugh 2012) also applied multi-scale filtering and a shape-based algorithm to improve the segmentation accuracy respectively. Due to the limited representation capability of the hand-crafted features, these methods are not able to deal with challenging cases and their performance also would be greatly degraded when the datasets become larger.

Recently, with the rise of deep learning, convolutional neural networks (CNN) have been used in this challenging task. Li et al. (Li et al. 2017) and Brandao et al. (Brandao et al. 2017) first used fully convolutional networks (FCNs) (Long, Shelhamer, and Darrell 2015) to segment the polyps. Sun et al. (Sun et al. 2019) introduced the dilated convolution (Yu and Koltun 2015) to improve the feature representation ability of U-Net (Ronneberger, Fischer, and Brox 2015). ResU-Net++(Jha et al. 2019) also employed atrous spatial pyramid pooling (Chen et al. 2017) to capture multi-scale contexts. Besides, Pra-Net(Fan et al. 2020) further proposed a reverse attention module to calibrate the misaligned predictions. However, these methods are still not able to efficiently bridge the semantic gap among different layers and utilize the non-local relations for better segmentation accuracy with real-time performance.

Semantic Gap within Multi-level Feature Maps

U-Net (Ronneberger, Fischer, and Brox 2015) and many its variants introduced skip connections to compensate for the loss of spatial details caused by multiple pooling operations. Despite preserving the dissipated spatial features, the shallow features are too noisy to provide sufficient high-resolution semantic guidance, resulting in a certain semantic gap in the fusion among different level features. A lot of effort have been devoted to alleviate this semantic gap. To guide the feature fusion with more semantic information, Exfuse (Zhang et al. 2018) embed high-level features into low-level features based on an element-wise multiplication. By introducing attention mechanisms, Attention U-Net (Oktay et al. 2018) further optimized this process. MultiResU-Net (Ibtehaz and Rahman 2020) integrated several convolutions into the skip connection to resolve the disparity within the features of different levels. SF-Net (Li et al. 2020) further improved the feature fusion between adjacent feature maps by minimizing semantic misalignment. In this paper, we propose two efficient semantic calibration and refinement modules to further address the semantic gap, which can take full advantage of the complementariness within multi-level feature maps to improve the segmentation performance.

Method

Network Architecture

The architecture of our proposed SCR-Net is as shown in Figure 2, which is mainly consists of a semantic calibra-

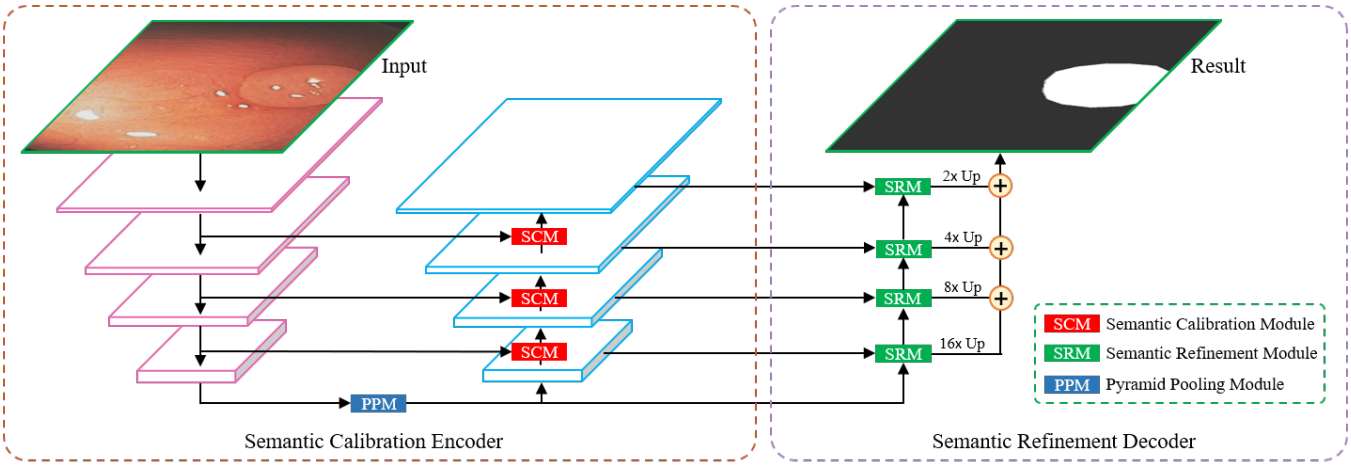

Figure 2: Our proposed SCR-Net, which is mainly consists of a semantic calibration encoder and a semantic refinement decoder. Our semantic calibration encoder obtains a better semantic transmission from deeper layer to the relative shallow layer based on semantic calibration, while our semantic refinement decoder also enhances the feature discrimination between the polyps and background tissues.

tion encoder and a semantic refinement decoder. As feature maps extracted in two neighboring layers contain different semantic information, we can utilize the higher-level semantic information in the deeper layer to enhance the feature map in current layer. To accurately transmit the rich higher-level semantic information from $(i+1)^{th}$ layer to i^{th} layer, we propose a semantic calibration module (SCM) to replace the traditional bilinear upsampling operation. By addressing the semantic misalignment problem based on our SCM, we can obtain a better semantic fusion between two neighboring feature maps in our encoder. For the top layer, we further employ a pyramid pooling module (PPM) (Zhao et al. 2017) to enrich the highest level semantics in our SCR-Net with global context information. On the other hand, we also introduce a semantic refinement module (SRM) to enhance the discrimination capability of the features for targeting objects in the decoder. According to the global context information, we can simultaneously strengthen the targets and weaken the backgrounds by re-weighting feature maps before the feature fusion in our decoder. Based on better distinction between the polyps and other tissues, our SCR-Net finally achieves a higher accuracy of polyp segmentation from colonoscopy videos.

Semantic Calibration Module

To reinforce the higher-level semantic information in each layer, we can employ a feature pyramid network (FPN) to rearrange the feature maps extracted in the original encoder, as shown in the middle of Figure 2. Because the resolutions of feature maps in two neighboring layers are different, we need upsample the higher-level feature map in $(i+1)^{th}$ layer to meet the resolution of feature map in i^{th} layer for feature pyramid fusion. Previously, the upsampling for traditional feature pyramid fusion is usually based on a simple bilinear upsampling, which may cause semantic misalignments and damage the transmission accuracy of higher-level semantics from deeper layer to shallow layer. In this paper, we employ a semantic calibration module (SCM) to replace the traditional bilinear upsampling operation, aiming to address the semantic misalignment problem based on the semantic offset calculated between two neighboring layers in the encoder.

The detailed implementation of our SCM is as shown in Figure 3. To achieve semantic calibration before feature pyramid fusion, The key issue for our SCM is how to accurately obtain the semantic offset between two neighboring layers. Specifically, given two adjacent feature maps $\mathbf{X}_{i+1} \in \mathbb{R}^{H\times W\times C}$ and $\mathbf{X}_i \in \mathbb{R}^{2H\times 2W\times C/2}$, where C is the number of channels and H and W is the height and width of the feature map, respectively. As shown in Figure 3, to obtain a feature map with the same resolution and channel number with $\mathbf{X}_i$, we first apply a 1×1 convolution and a bilinear upsampling to $\mathbf{X}_{i+1}$, which can be written as

$$\mathbf{X}^{'}_{i+1} = \mathbf{Up}(\mathbf{Conv}_{1\times 1}(\mathbf{X}_{i+1})) \tag{1}$$

where $\mathbf{Conv}_{1\times 1}(\cdot)$ denotes the 1×1 convolution, and the $\mathbf{Up}(\cdot)$ denotes the bilinear upsampling. After concatenating $\mathbf{X}^{'}_{i+1}$ and $\mathbf{X}_i$ together, we can further obtain a semantic offset $\mathbf{S}_{i+1}$ with 2 channels based on a 3×3 convolution. Therefore, our semantic offset $\mathbf{S_{i+1}}$ can be written as

$$\mathbf{S}_{i+1} = \mathbf{Conv}_{3\times 3}(\mathbf{Cat}(\mathbf{X}^{'}_{i+1}, \mathbf{X}_i)) \tag{2}$$

where $\mathbf{Conv}_{3\times 3}(\cdot)$ denotes the 3×3 convolution, and the $\mathbf{Cat}(\cdot)$ represents a concatenation along the channel dimension. Note that, our 3×3 convolution calculates the semantic offset based on the corresponding semantic flow field between $\mathbf{X}^{'}_{i+1}$ and $\mathbf{X}_i$. Based on the semantic offset $\mathbf{S}_{i+1}$, we can calibrate the semantics in the higher-level feature map $\mathbf{X}_{i+1}$ by remapping each pixel in $\mathbf{X}_{i+1}$ to a grid with the same size as $\mathbf{X}_i$. Finally, the calibrated higher-level feature map can be fused with the feature map $\mathbf{X}_i$, through which we can accurately transmit the higher-level semantics from

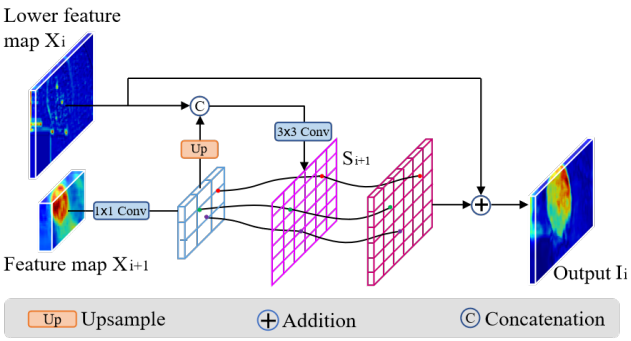

Figure 3: Semantic calibration module. By calibrating the feature pyramid fusion based on the semantic flow offset, our SCM can obtain a better transmission from the rich higher-level semantic information in $(i+1)^{th}$ layer to the adjacent i^{th} layer.

deeper $(i+1)^{th}$ layer to i^{th} layer for our feature pyramid fusion in the encoder. Mathematically, above process can be written as

$$\mathbf{I_i} = \mathbf{X_i} \oplus \mathbf{Map}(\mathbf{X'_{i+1}}, \mathbf{S_{i+1}}) \tag{3}$$

where the $\oplus$ represents the broadcast element-wise addition, the $\mathbf{I}_i$ represent the output of the SCM in the $(i+1^{th}$ layer, and the $\mathbf{Map}$ is implemented by adding a corresponding semantic offset to each pixel in the feature map. Obviously, our SCM addresses the semantic misalignment problem between two neighboring layers. Based on the semantic calibration for the higher-level feature map, our SCR-Net obtain a better semantic transmission from deeper layer to the relative shallow layer in feature pyramid fusion, which improves the accuracy of polyp segmentation.

Semantic Refinement Module

Based on semantic reinforced feature map in each layer in our semantic calibration encoder, we need further fuse multi-level feature maps and restore the final dense prediction to the original spatial resolution. However, due to the contrast between the target objects and the backgrounds can be very low for our polyp segmentation task, which makes the target features and the background features very difficult to discriminate, especially for the challenging cases where have blur polyp boundaries, as shown in Figure 1. To enhance the discrimination capability of the features between the targeting polyps and other background tissues, we introduce a semantic refinement module (SRM) to simultaneously strengthen the targets and weaken the backgrounds by re-weighting features before the fusion between two adjacent feature maps in our decoder.

As depicted in Figure 4, our proposed SRM takes two feature maps $\mathbf{I}_i \in \mathbb{R}^{C\times H\times W}$ and $\mathbf{I}_{i+1} \in \mathbb{R}^{2C\times H/2\times W/2}$ as input, where $\mathbf{I}_i$ and $\mathbf{I}_{i+1}$ are two neighboring feature maps produced by our semantic calibration encoder. Specifically, the input feature map $\mathbf{I}_{i+1}$ is first mapped to the feature $\mathbf{U_v} = [U_{v_1}, U_{v_2}, ..., U_{v_c}] \in \mathbb{R}^{C\times 1\times 1}$ by a global average pooling (GAP) (Lin, Chen, and Yan 2013) for global context

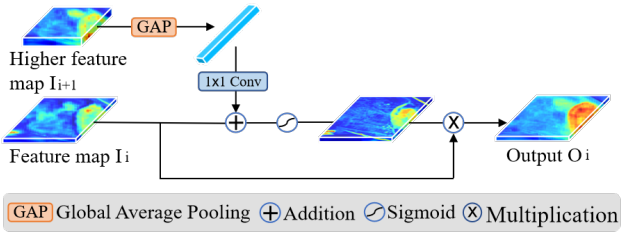

Figure 4: Semantic refinement module. To enhance the discrimination capability of the features between the targeting polyps and other background tissues, we introduce a semantic refinement module (SRM) to simultaneously strengthen the targets and weaken the backgrounds by re-weighting features based on global context information.

modeling. The U_{v_i} is the i^{th} element of $\mathbf{U_v}$, which can be calculated based on the i^{th} channel descriptor of the feature $\mathbf{I}_{i+1}$ and written as

$$U_{v_i} = \frac{1}{H\times W}\sum_{j=1}^{H}\sum_{k=1}^{W} f_i(j,k) \tag{4}$$

where $f_i(j,k)$ indicates a local pixel value at position (j,k) in the i^{th} channel of feature $\mathbf{I}_{i+1}$. To align $\mathbf{U_v}$ and lower-level feature map $\mathbf{I}_i$ with the same channel number, we transform the feature map $\mathbf{U_v}$ using a channel rescaling function ϕ_v, which can be written as

$$\hat{\mathbf{U}}_\mathbf{v} = \phi_v(\mathbf{U_v}, \theta) \tag{5}$$

where $\hat{\mathbf{U}}_\mathbf{v}$ is a unified feature vector. θ is the relevant learnable parameter, and the ϕ_v is simply implemented through a 1×1 convolution followed by a ReLU activation function. After that, the global context information is embedded into each pixel in $\mathbf{I}_i$ by broadcast element-wise addition for feature fusion. In this way, the target features and background features in the lower-level feature map $\mathbf{I}_i$ can be associated with global context information, so we can further employ a gating function to generate a global-relation weighting map $\hat{\mathbf{U}}_g$, which can be written as

$$\hat{\mathbf{U}}_g = \delta(\hat{\mathbf{U}}_\mathbf{v} + \mathbf{I}_i) \tag{6}$$

where δ refers to the Sigmoid function. Finally, we get the output feature map $\mathbf{O}_i$ by multiplying $\mathbf{I}_i$ with the global-relation weight map $\hat{\mathbf{U}}_g$.

$$\mathbf{O}_i = \mathbf{I}_i \otimes \hat{\mathbf{U}}_g \tag{7}$$

where $\otimes$ is broadcast element-wise multiplication. Therefore, we can obtain re-weighted feature maps by strengthening the distinction between the target features and background features. By enhancing feature discrimination between polyps and the background tissues, our semantic refinement module can further improve the segmentation accuracy.

Loss Function

For the challenging polyp segmentation task, we train our framework with the designed loss function which composed of a cross entropy loss and a dice coefficient loss (Crum, Camara, and Hill 2006) (Milletari, Navab, and Ahmadi 2016), as shown in Eq 8. Cross entropy loss function is a well-validated classification loss, which is widely used in pixel-wise classification problems. However, if we only apply a single cross entropy loss, pixels can be easily classified to the classes with more samples due to the class imbalance. Considering that our polyp segmentation task has the problem of class imbalance, where some polyps appear with in the very small regions, we also equip a dice coefficient loss to our loss function. Therefore, to further address the problem of class imbalance problem, our loss function can be written as

$$\textbf{Loss} = \lambda C + (1-\lambda)D + \gamma\|\omega\|_2^2, \lambda \in [0,1] \qquad (8)$$

where C indicates the binary cross entropy loss. D is the dice coefficient loss, and $\gamma\|\omega\|_2^2$ represents the L_2 regularization loss (Hoerl and Kennard 1970) used to avoid overfitting. We also balance the C and D by a weight coefficient λ. In our experiments, the loss has the best performance when we set the λ to 0.4. For a fair comparison, we apply the same loss function in all experiments, including the following comparisons with state-of-the-art methods.

Experiments

Dataset and Evaluation Metrics

In our experiments, we used the Kvasir-SEG dataset to evaluate the performance of our proposed SCR-Net. The Kvasir-SEG dataset is obtained from colonoscopy videos, which contains 1000 frames with a resolution ranging from 332×487 to 1920×1072 pixels. Since the Kvasir-SEG dataset does not provide an additional test dataset, we randomly divided the 1000 images into 700 images for training, 100 images for validation and 200 images for testing. To normalize the image resolution and improve the computational efficiency, we also resized all images to 448×448 pixels.

To measure the performance of polyp segmentation, we adopt the four commonly-used metrics to evaluate our SCR-Net and other competitors on this dataset, including intersection over union (IoU), dice coefficient (Dice), sensitivity (SE), specificity (SP).

Implementation Details

We implemented our SCR-Net with PyTorch on a single NVIDIA RTX2080Ti GPU card with 11 GB memory. To accelerate the convergence in training phase, we employed Kaiming initialization (He et al. 2015) to initialize the parameters of our model. To avoid overfitting, we also performed 6 kinds of data augmentations to increase the numbers of the images, including horizontal flipping, vertical flipping, random contrast adjustment, zoom out and zoom in with a coefficient of 0.5 and 2.0, and random cropping. The batch size is empirically set to 8. To obtain a fast convergence, we also employ the Adam optimizer to train our model, where the initial learning rate is set to 0.001. To further achieve better model training, we also adopt the Gradual Warmup (Goyal et al. 2017) in fine tuning the decay learning rate. Specifically, we first setup a small learning rate lr_{min} and gradually increase it to the initial learning rate. This step can be expressed as $lr = lr_{min} \times (1 + \frac{step}{warmup\_step})^{power}$, where lr_{min} and $power$ are set to 0.00001 and 0.9, respectively. Then the learning rate gradually decreases by $lr = lr_{init} \times (1 - \frac{step}{total\_step - warmup\_step})^{power}$, where $power$ is set to 0.9, and $total\_step = epoch \times \frac{num\_samples}{batch\_size}$. As the learning rate decays, the models can gradually achieve the approximate global optimum.

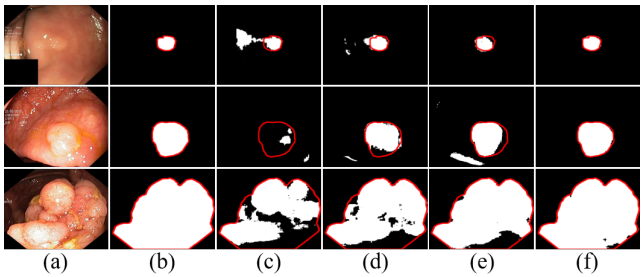

Figure 5: Visual comparison of our ablation study for SCM and SRM. Ground truth is outlined with red lines. (a) Original image. (b) Ground truth. (c) Baseline. (d) Baseline+SCM. (e) Baseline+SRM. (f) Ours (Baseline+SCM+SRM).

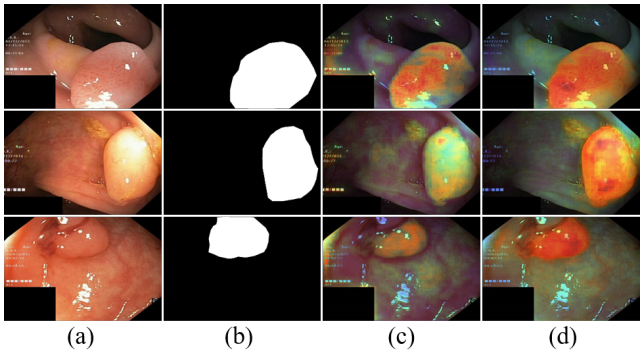

Figure 6: Visual comparison of feature maps in our ablation study for SRM. (a) Original image. (b) Ground truth. (c) Feature map extracted without SRM. (d) Feature map with SRM.

Ablation Studies

To demonstrate the performance of our proposed SCR-Net, we conducted the following ablation studies to evaluate the effectiveness of SCM and SRM based on the Kvasir-SEG dataset. In our experiments, we used a light-weight feature pyramid network (FPN) equipped with PPM as our Baseline.

Ablation Study for SCM As mentioned before, our SCM is inserted to the FPN framework to address the semantic misalignment problem in transmitting the higher-level se-

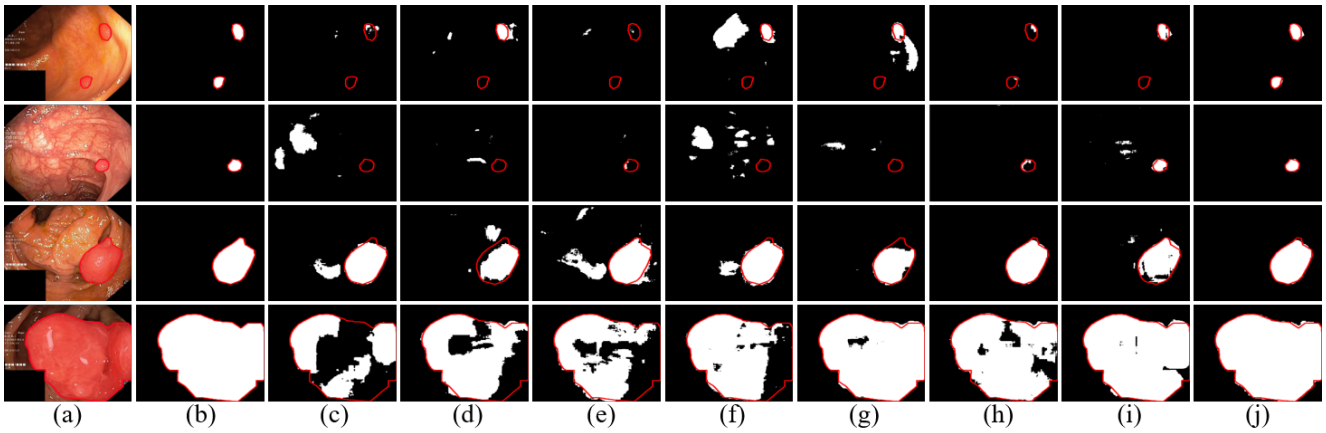

Figure 7: Visual comparison with state-of-the-art methods. (a) Input images. (b) Ground truth. (c) U-Net. (d) PSP-Net. (e) Attention U-Net. (f) ResUNet++. (g) CE-Net. (h) CPF-Net. (i) Pra-Net. (j) Ours.

Method	IoU (%)	SE (%)	SP (%)
Baseline (FPN+PPM)	64.96	84.13	94.98
Baseline+SCM	68.48	87.25	96.89
Baseline+SRM	67.43	86.43	96.28
Baseline+SCM+SRM	**70.85**	**89.17**	**98.21**

Table 1: Statistical comparison of our ablation study for SCM and SRM.

Method	Year	IoU(%)	Dice(%)	SE(%)	SP(%)	FPS
U-Net	2015	63.89	75.18	83.24	94.78	31
PSP-Net	2017	66.27	75.92	85.27	95.39	27
Attention U-Net	2018	65.93	75.46	84.92	95.14	23
ResUNet++	2019	66.42	76.12	86.14	95.97	9
CE-Net	2019	67.28	76.97	87.14	96.36	24
CPF-Net	2020	69.43	78.39	86.81	96.83	21
Pra-Net	2020	69.07	78.56	87.37	96.53	23
Ours	**2020**	**70.85**	**80.78**	**89.17**	**98.21**	**32**

Table 2: Statistical comparison of with different state-of-the-art methods.

mantic information from deeper layers to their adjacent layers. Note that, we also have applied a PPM (Zhao et al. 2017) to enrich the highest level semantics in the top layer of our semantic calibration encoder, as shown in Figure 2.

Typical challenging cases for visual comparison of our ablation study are as shown in Figure 5, where ground truth is outlined with red lines. Without addressing the semantic misalignment problem in sematic transmission, Baseline method still cannot obtain accurate polyp segmentation, as shown in Figure 5 (c). After calibrating the semantics, we can observe much better segmentation results in Figure 5 (d), clearly indicating the effectiveness of our SCM in handing the multi-scale polyp segmentation.

In addition, we also performed a statistical comparison in our ablation studies by collecting the mean IoU, SE and SP values over the Kvasir-SEG dataset. As shown in Table 1, the Baseline+SCM method achieves 68.48%, 87.25%, 96.89% in terms of IoU, SE and SP metrics, which outperforms the Baseline by 3.52%, 3.12% and 1.91%, respectively. This also implies that our SCM can effectively improve the polyp segmentation by utilizing the better higher-level sematic transmission between two neighboring layers based on the proposed semantics calibration.

Ablation Study for SRM By fully utilizing the global context information, our SRM enhances feature discrimination between target objects and background tissues. Compared with the Baseline method, which may suffer serious ambiguity between the target objects and background tissues, our Baseline+SRM (Figure 5 (e)) can obtain much better polyp segmentation results, especially for the boundary regions of the target objects, clearly indicating the effectiveness of our SRM in handing low contrast between polyps and the background tissues in the challenging polyp segmentation. On the other hand, the statistical results shown in Table 1 also verify the advantages of our SRM,, where Baseline+SRM achieves 67.43%, 86.43%, 96.28% in terms of IoU, SE and SP metrics, outperforming the Baseline by 2.47%, 2.30% and 1.30%, respectively.

In addition, we also visualized the feature maps to further investigate the effectiveness of SRM. As shown in Figure 6, without the guidance of SRM, Baseline (Figure 6 (c)) cannot easily distinguish the boundaries between the target objects and background regions. In contrast, feature maps obtained with SRM clearly validates that we can achieve much better discrimination between target objects and background tissues, with much clear boundaries as shown in Figure 6 (d).

Finally, by seamless integrating both SCM and SRM in our SCR-Net, we cannot only obtain better higher-level sematic transmission between two neighboring layers in encoder, but also enhances feature discrimination between target objects and background tissues in the decoder, as shown in the Figure 5 (f).

Comparison with the State-of-the-art Methods

To further verify the segmentation performance of proposed SCR-Net, we also compared our method with several state-of-the-art methods, including U-Net (Ronneberger, Fischer, and Brox 2015), PSP-Net (Zhao et al. 2017), Attention U-Net (Oktay et al. 2018), ResUNet++ (Jha et al. 2019), CE-Net (Gu et al. 2019), CPF-Net (Feng et al. 2020), and Pra-Net (Fan et al. 2020). To guarantee a fair comparison, all of approaches are implemented under the same computing environments and with the same data augmentations. Note that all methods are trained from scratch without loading any pre-trained weights.

Visual comparison for different competitors on several challenging cases is as shown in Fig7, where we can observe the major challenging issues for polyp segmentation, including irregular shapes, scale variations, and low contrast between the polyps and the surrounding tissues. By simply stacking the continuous convolution and pooling operations, U-Net still cannot accurately segment some challenging cases. Based on region-based context aggregation to exploit global contexts, PSP-Net obtains better segmentation results than U-Net. On the other hand, by capturing more high-level information and retaining rich spatial information, CE-Net also obtains improvements in polyp segmentation. Similarly, ResUNet++ also achieves improvements in segmenting polyps with different scales based on multi-scale feature extraction. However, above methods still cannot tackle the issues of blurred boundaries and low contrast between target objects and background tissues, especially where have very much irrelevant background noises. Attention U-Net tried to use a novel attention gate to suppress irrelevant background noise while highlighting the target region, but it is still cannot handle well the blurred boundary discrimination, especially for the small targets, as shown in the first and second rows in the Figure 7. More recently, by exploiting and fusing rich contexts progressively through a global pyramid guidance module and a scale-aware pyramid fusion module, CPF-Net can achieve better segmentation results than the above methods. To address the blur boundary between the polyps and its surrounding tissues, Pra-Net further develops a reverse attention module to calibrate misaligned predictions. Without fully utilizing multi-level features based on an excellent sematic transmission between different levels of feature maps, where feature discrimination ability is also weak between target objects and background tissues, all the above competitors are still not obtain satisfactory segmentation results for polyp segmentation. Based on a seamless combination of both SCM and SRM, our SCR-Net generally outperform existing competitors. As shown in Figure 7 (j), compared with other methods, our results are the closest to the ground truth, which cannot only handle well the targets with large variations of sizes, but also better deal with the target objects with low contrast between polyps and background tissues.

In addition, we also performed a statistical comparison by collecting the mean IoU, Dice, SE, SP, and FPS values. Our proposed method achieves scores of 70.85%, 80.78%, 89.17%, and 98.21% in terms of IoU, Dice, SE, and SP metrics respectively, which also generally outperforms other state-or-the-art methods. Compared with the widely used U-Net, our SCR-Net improves IoU and Dice by approximate 7.0% and 5.6%, respectively. More importantly, our SCR-Net can still reach a speed of 32 FPS on a commonly-used GPU card (NVIDIA GTX 965M), which is sufficient to be applied to colonoscopy procedures for real-time segmentation of polyps.

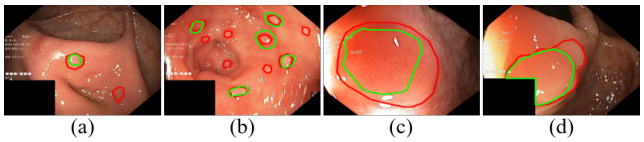

Figure 8: Failure cases. Green and red contours denote our segmented polyps and the ground truth, respectively.

Discussions and Limitations

Through the above extensive experiments, it is clearly observed that our SCR-Net achieves the best performance and successfully solves the major polyp segmentation challenges, including irregular shapes, scale variations, and low contrast between the polyps and the surrounding tissues. By addressing the semantic misalignment problem based on our SCM, we obtain a better semantic fusion between two neighboring feature maps in our encoder, which also provides a good tool for segmenting objects with multi-scale problem or with multi-level feature fusion. By simultaneously strengthening the targets and weakening the backgrounds with a re-weighting strategy based on global context, we also enhance the discrimination capability of the features for targeting objects in the decoder, which also provides a good hint to solve similar segmentation tasks with blur boundaries among different classes of targets.

Although ablation studies and comparisons have demonstrated the effectiveness of our SCR-Net, our method stil l has some limitations. As shown in Figure 8, our method still may be failed when there contains multiple extremely small polyps, such as the cases shown in Figure 8 (a)-(b), or the color contrast between polyps and the background is extremely low, such as the cases shown in Figure 8 (c)-(d). However, as our SCR-Net can still handle well most of polyp segmentation except the extreme cases, which remains potentially a useful tool for a real-time polyp detection and operation system.

Conclusion

We present a novel ConvNet with two new semantic calibration and refinement techniques to bridge the semantic gap between feature maps at different levels for more accurate polyp segmentation. More important, the proposed techniques are efficient enough to maintain real-time performance when conducting segmentation, which is critical in clinical practice. Extensive experiments demonstrate the effectiveness of the proposed ConvNet. Future investigations include testing it on more datesets and integrating it into colonoscopy procedures.

Acknowledgements

This work was supported in part by grants from the National Natural Science Foundation of China under Grant 61973221, the Natural Science Foundation of Guangdong Province of China under Grant 2018A030313381 and Grant 2019A1515011165, the Key Lab of Shenzhen Research Foundation of China under Grant 201707311550233, the COVID-19 Prevention Project of Guangdong Province of China under Grant 2020KZDZX1174, the Major Project of the New Generation of Artificial Intelligence of China under Grant 2018AAA0102900, and the Hong Kong Research Grants Council of China under Grant PolyU 152035/17E.

References

Brandao, P.; Mazomenos, E.; Ciuti, G.; Caliò, R.; Bianchi, F.; Menciassi, A.; Dario, P.; Koulaouzidis, A.; Arezzo, A.; and Stoyanov, D. 2017. Fully convolutional neural networks for polyp segmentation in colonoscopy. In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, 101340F.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40(4): 834–848.

Crum, W. R.; Camara, O.; and Hill, D. L. 2006. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE transactions on medical imaging* 25(11): 1451–1461.

Fan, D.-P.; Ji, G.-P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; and Shao, L. 2020. Pranet: Parallel reverse attention network for polyp segmentation. *arXiv preprint arXiv:2006.11392* .

Feng, S.; Zhao, H.; Shi, F.; Cheng, X.; Wang, M.; Ma, Y.; Xiang, D.; Zhu, W.; and Chen, X. 2020. CPFNet: Context Pyramid Fusion Network for Medical Image Segmentation. *IEEE Transactions on Medical Imaging* .

Ganz, M.; Yang, X.; and Slabaugh, G. 2012. Automatic segmentation of polyps in colonoscopic narrow-band imaging data. *IEEE Transactions on Biomedical Engineering* 59(8): 2144–2151.

Goyal, P.; Dollár, P.; Girshick, R.; Noordhuis, P.; Wesolowski, L.; Kyrola, A.; Tulloch, A.; Jia, Y.; and He, K. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677* .

Gross, S.; Kennel, M.; Stehle, T.; Wulff, J.; Tischendorf, J.; Trautwein, C.; and Aach, T. 2009. Polyp segmentation in NBI colonoscopy. In *Bildverarbeitung für die Medizin 2009*, 252–256.

Gu, Z.; Cheng, J.; Fu, H.; Zhou, K.; Hao, H.; Zhao, Y.; Zhang, T.; Gao, S.; and Liu, J. 2019. CE-Net: context encoder network for 2D medical image segmentation. *IEEE transactions on medical imaging* 38(10): 2281–2292.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, 1026–1034.

Hoerl, A. E.; and Kennard, R. W. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12(1): 55–67.

Ibtehaz, N.; and Rahman, M. S. 2020. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Networks* 121: 74–87.

Jerebko, A.; Franaszek, M.; and Summers, R. 2002. Radon transform based polyp segmentation method for CT colonography computer aided diagnosis. In *RADIOLOGY*, volume 225, 257–258.

Jerebko, A. K.; Teerlink, S.; Franaszek, M.; and Summers, R. M. 2003. Polyp segmentation method for CT colonography computer-aided detection. In *Medical imaging 2003: physiology and function: methods, systems, and applications*, volume 5031, 359–369.

Jha, D.; Smedsrud, P. H.; Riegler, M. A.; Halvorsen, P.; de Lange, T.; Johansen, D.; and Johansen, H. D. 2020. Kvasir-seg: A segmented polyp dataset. In *International Conference on Multimedia Modeling*, 451–462.

Jha, D.; Smedsrud, P. H.; Riegler, M. A.; Johansen, D.; De Lange, T.; Halvorsen, P.; and Johansen, H. D. 2019. ResUNet++: An Advanced Architecture for Medical Image Segmentation. In *2019 IEEE International Symposium on Multimedia (ISM)*, 225–2255.

Kolligs, F. T. 2016. Diagnostics and epidemiology of colorectal cancer. *Visceral medicine* 32(3): 158–164.

Leufkens, A.; Van Oijen, M.; Vleggaar, F.; and Siersema, P. 2012. Factors influencing the miss rate of polyps in a back-to-back colonoscopy study. *Endoscopy* 44(05): 470–475.

Li, Q.; Yang, G.; Chen, Z.; Huang, B.; Chen, L.; Xu, D.; Zhou, X.; Zhong, S.; Zhang, H.; and Wang, T. 2017. Colorectal polyp segmentation using a fully convolutional neural network. In *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 1–5.

Li, X.; You, A.; Zhu, Z.; Zhao, H.; Yang, M.; Yang, K.; Tan, S.; and Tong, Y. 2020. Semantic Flow for Fast and Accurate Scene Parsing. In *European Conference on Computer Vision*, 775–793. Springer.

Lin, M.; Chen, Q.; and Yan, S. 2013. Network in network. *arXiv preprint arXiv:1312.4400* .

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, 565–571.

Nappi, J.; and Yoshida, H. 2002. Automated detection of polyps with CT colonography: evaluation of volumetric features for reduction of false-positive findings. *Academic Radiology* 9(4): 386–397.

Oktay, O.; Schlemper, J.; Folgoc, L. L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N. Y.; Kainz, B.; et al. 2018. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999* .

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241.

Sun, X.; Zhang, P.; Wang, D.; Cao, Y.; and Liu, B. 2019. Colorectal Polyp Segmentation by U-Net with Dilation Convolution. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 851–858.

Yao, J.; Miller, M.; Franaszek, M.; and Summers, R. M. 2004. Colonic polyp segmentation in CT colonography-based on fuzzy clustering and deformable models. *IEEE Transactions on Medical Imaging* 23(11): 1344–1352.

Yoshida, H.; Masutani, Y.; MacEneaney, P.; Rubin, D. T.; and Dachman, A. H. 2002. Computerized detection of colonic polyps at CT colonography on the basis of volumetric features: pilot study. *Radiology* 222(2): 327–336.

Yu, F.; and Koltun, V. 2015. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122* .

Zhang, Z.; Zhang, X.; Peng, C.; Xue, X.; and Sun, J. 2018. ExFuse: Enhancing Feature Fusion for Semantic Segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.