Polyp Segmentation of Colonoscopy Images by Exploring the Uncertain Areas

QINGQING GUO\textsuperscript{1}, XIANYONG FANG\textsuperscript{1}, (Member, IEEE), LINBO WANG\textsuperscript{1}, AND ENMING ZHANG\textsuperscript{2}

\textsuperscript{1}School of Computer Science and Technology, Anhui University, Heifei 230601, China
\textsuperscript{2}Islet Pathophysiology, Department of Clinical Science, Lund University Diabetes Centre, Malmo 21428, Sweden

Corresponding author: Xianyong Fang (e-mail: fangxianyong@ahu.edu.cn).

This work is supported by the the Natural Science Foundation of Anhui Province, China (Grant No. 2108085MF210).

\section*{ABSTRACT} Colorectal cancer is one of the leading causes of death worldwide. Polyps are early symptoms of colorectal cancer and prone to malignant transformation. Polyp segmentation of colonoscopy images can help diagnosis. However, existing studies on polyp segmentation of colonoscopy images face two main difficulties: blurry polyp boundaries, close resemblances between polyps and surrounding tissues. The former may lead to partial segmentations, while the latter can result in false positive segmentations. This paper proposes a new polyp segmentation framework to tackle the two challenges. In this method, an uncertainty region based module called Uncertainty eXploration (UnX) is introduced to get the integral polyp region while eliminating the interferences from the backgrounds. Specifically, it refines the feature maps with ternary guidance masks by dividing the initial guidance maps into three types: foreground, background and uncertain region, so that the uncertain areas are highlighted for more foreground objects while the backgrounds are forcefully suppressed to avoid interferences of tissues in background. Taking UnX as side supervision to the transformer encoder based backbone stages, the proposed method can mine the boundary areas from the uncertainty regions gradually and obtain integral polyp segmentation finally. Moreover, a new module called Feature Enhancement (FeE) is also incorporated in the framework to enhance the discrimination for images with significant variation of sizes and shapes of polyps. FeE can supply multi-scale features to the global oriented transformer features. Experiments on five polyp segmentation benchmark datasets of colonoscopy images, Kvasir, CVC-ClinicDB, ETIS, CVC-ColonDB and CVC-300, show the superior performances of our proposed method. Especially, for ETIS, the most challenging among the five datasets, our method achieves 7.7\% and 5.6\% improvements in mDSC and mIoU respectively in comparison with the state-of-the-arts methods.

\section*{INDEX TERMS} Polyp segmentation, Deep learning, Colonoscopy image, Colorectal cancer

\section*{I. INTRODUCTION} COLORECTAL cancer (CRC) is serious for human health, leading to the fourth highest Cancer mortality rate in the world [1]. Polyps in the intestinal tissues are precursors of CRC and can easily turn into malignant lesions [2]. Colonoscopy can provide information on the location and appearance of colorectal polyps, allowing doctors to remove them before turning into cancer. However, colonoscopy can sometimes miss the detection for polyp. Therefore, automatic and accurate polyp segmentation of colonoscopy images is of great significance, which can provide additional support to clinicians.

Nowadays, deep learning based polyp segmentation methods [3]–[12] are popular and demonstrated to be more effective than the traditionally hand-crafted feature based methods [13]–[15]. Popular methods take U-shape structures [16], [17] and their variants [6], [10], [18]–[20]. Recently, the attention mechanism is often adopted to either restore clear boundaries between polyps and surrounding tissues [3], [4] or extract global and local features [21]–[24].

These methods are still not robust enough to cope with the challenging cases where there are blurry boundaries (e.g., the polyp in the top image of Fig. 1) and close resemblances between polyps and surrounding tissue in colonoscopy images (e.g., the tissue in bottom image of Fig. 1(a)). Consequently, partial segmentations of polyps and false positive segmenta-
This uncertainty exploration based idea can also be explained from the human cognition [32]: Humans has a deliberation ability that they can keep seeking the interesting parts of an object without the interference of background, after initially seeing its salient parts. However, most of the existing studies [8], [21], [33] don’t consider the explicit background priors, which may easily obtain false positive segmentation because of the similarities between polyps and surrounding tissues [34]. Recent state-of-the-arts methods try the direct foreground/background binary division based background priors [3], [4]. However, this binary division can only get a rough foreground prior with the possible inclusion of fake foreground objects in the background. Therefore, it also cannot avoid false positive segmentations and partial segmentations. This paper, however, explores the uncertain areas after separating the whole feature map into three areas with a ternary mask and, therefore, explicitly provides such a background prior to help the network fulfill the robust segmentation process.

It is also worthwhile to note that, in the weakly supervised study for nature image segmentation, Hou et al. [35] ever proposed a ternary mask based method, where both ternary and binary masks are applied once sequentially. They need the binary mask for finding more foreground areas from the background because of the possible errors, i.e., missing foreground objects, brought by the ternary mask due to their weak image-level supervision. However, our method takes a strong supervised way so that the ternary-mask is strong enough to help obtaining salient foregrounds and backgrounds. In addition, multiple ternary-mask based side supervisions can gradually to mine the boundary areas for integral polyps segmentations.

Our main contributions can be summarized as follows:

- An uncertainty exploration module, UnX, which take a ternary guidance mask to explore the uncertain areas for more foreground objects while suppressing the interference of the background, and thus boost the discrimination of the network for a robust foreground recognition.
- A feature enhancement module, FeE, which takes parallel atrous convolutions to augment the multi-scale feature representation with finer scales and thus helps improve polyp segmentation under different scales.
- A deep polyp segmentation framework for colonoscopy images, which integrates the two proposed modules, UnX and FeE, into the transformer encoder for more accurate segmentations of polyp images.

II. RELATED WORK

Existing methods can be classified into two types: traditional methods and deep learning based methods. The former adopts artificially designed features while the latter utilizes the neural network to automatically extract them. Traditional methods are mainly based on low-level features, such as textures [14], geometric features [14] or superpixels [36]. However, they tend to have poor segmentation performances, especially in comparison with the deep learning based meth-
ods. Therefore, our main focus here is on the deep learning based methods.

Many deep learning based methods have been proposed for polyp segmentation. Some of them [37]–[40] directly use some basic deep learning models for polyp segmentation. For example, Vazquez et al. [38] adopted a full convolutional network (FCN), which was perhaps the first deep learning work for polyp segmentation; Qadir et al. [40] employed Mask R-CNN for joint polyp detection and segmentation learning. Those models focus on the basic application of deep learning models and thus obtain limited performances, especially for the two main challenges.

Improved performances with more complex U-shape designs are also proposed [6], [17], [20], [30], [31], which is based on U-Net [16]. For example, Alam et al. [6] replaced convolution back in U-Net with ResNet block; Sun et al. [30] introduced a dilated convolution to learn high-level semantic features in the U-shape structure.

Considering that the boundary of a polyp is often blurry, some researchers [41]–[43] took extra boundaries as supervision, which however need more edge ground truth.

Attention mechanism is also adopted by some methods [3], [4], [33], where robust salient foreground regions provide important initial cues for segmentation. Among them, foreground/background binary division oriented methods [3], [4] achieve the state-of-the-arts performances. However, these methods build on the binary classification of the whole image into foreground and background without considering the uncertainty of background. Kim et al. [4] incorporated an additional uncertain area though a bias operation. But such uncertain areas are actually very thin and along the foreground edges and, therefore, can only get limited performance gain as edge supervision. We, however, propose the ternary mask based UnX to uncover more foreground areas from much wider uncertain regions, which acts as side-supervision to fully explore the uncertain areas and suppress the interference from the apparent background objects.

The recent development of vision transformer spawns novel attention based methods [21]–[24], [44]. For example, Transfuse [21] and SwinE-Net [44] employs a two-branch architecture combining CNNs and transformers in a parallel style, while Dong et al. [22] applied the transformer to the backbone. These latest ideas show impressive results with higher accuracies. But transformers can only extract global features. Our method additionally integrates FeE to the transformer encoder for more fine scales and takes UnX to discover more foreground objects from the uncertain regions.

There are some studies on learning robust multi-scale features [10], [19], [30], [31], [42] for processing the vast varieties of polyp shapes and sizes. For example, some studies [10], [42] adopt standard convolution or pooling of different sizes to extract multi-scale features. However, standard convolution may require more tuning parameters or have limited receptive field, and pooling may cause loss of spatial information. Luckily, atrous convolution [45] was put forward to enlarge the receptive field, which was adopted by some studies [19], [30], [31] to do polyp segmentation. For example, Sun et al. [30] adopted only one atrous convolution at the end of encoder to widen the receptive field, but it is difficult to capture the appropriate context with only one atrous convolution. ResUNet++ [31] introduces atrous spatial pyramid pooling (ASPP) [46] to capture multi-scale context information, which adopts larger atrous rates, but it can be insufficient to deal with small polyps. Our proposed FeE adopts several atrous convolutions in parallel with different atrous rates appropriate for polyps, which can extract multi-scale context information for large size variation of polyps.

### III. THE PROPOSED METHOD

This section will first introduce the structure of the proposed framework briefly and then present two main modules, UnX and FeE, in details.

Overall, the proposed method targets at segmenting polyps in colonoscopy images robustly without blurry boundaries and the distraction of background objects. Its main strategy is to explore the uncertain areas through a ternary mask to remove the apparent backgrounds, which also can be taken as mimicking the human deliberation process with an explicit background prior.

The first important component is the salient foreground extraction which provides important foreground cues for creating an effective ternary guidance mask. Here, the transformer [26], [27], [47] which has been proved to exhibit stronger performances than the traditional CNN methods [48], [49] is adopted. In particular, the transformer encoder is applied so that strong backbone features can be obtained to support the robust extraction of the salient foreground parts and further complete and efficient segmentation.

Then the features from different stages of the encoder can be merged through a Partial Decoder (PD) module [28] to obtain the segmentation result. However, this backbone can only detect salient foreground regions and thus be difficult to localize the whole objects, especially when there are blurry boundaries between polyps and surrounding tissues. Therefore, UnX is integrated, which takes the ternary guidance mask to expand the foreground areas from the uncertain regions. UnX relies on the features from PD, which are generated with those from the encoder transformers.

However, the transformer features can only provide limited information about the global distribution [21] and thus miss the important local details for robust segmentation of size-and-shape varied polyps. Therefore, FeE is integrated at various stages and thus supplies rich features in both local and global scales for boosting the performance of UnX.

Fig. 2 demonstrates the principle of the proposed method with the refined features progressively obtained by UnX and FeE. It can be seen that UnX can help obtaining more complete foreground features than the baseline (Fig. 2(d)). However, its features are less rich and discriminative than those obtained after additionally integrating FeE (Fig. 2(e)).

Accordingly, our framework is designed in this way (Fig. 3). It takes the four-stage transformer encoder T as
Wang et al. [27] to obtain the initial backbone features and adopts FeE to each encoder stage so that multi-scale features can be obtained as input to PD. FeE is not considered for the first backbone stage because the feature is too coarse. The PD features $U^t$ are then input to UnX which combines them with the backbone features $F^{(5)}$ from the last encoder stage to enhance the discrimination of the network, so that more foreground objects in the uncertain areas can be discovered. The enhanced features $U^{(4)}$ are then up-sampled and input to the next UnX. There they will be combined again with the corresponding backbone features $F^{(4)}$ in the higher scales to further boost foreground features in the uncertain areas. This process repeats at the previous encoder stages with features $F^{(3)}$ and $F^{(2)}$ consequently till reaching the coarsest layer which is too coarse to consider. There will be correspondingly enhanced features $U^{(3)}$ and $U^{(2)}$ from UnX modules and the final features $U^{(2)}$ is taken as the prediction after applying the sigmoid function.

Formally, the whole process can be written as

$$F^{(i)} = T(I) \quad (i = 1, 2, 3, 4),$$

$$P = U(F^{(2)}, U(F^{(3)}, U(F^{(4)}, U^{(5)}))),$$

where: 1) $I$ represents the input image; 2) $U$ and $P$ represent the functions of UnX and FeE respectively; and 3)

$$U^{(5)} = P(F^{(5)}),$$

where $P$ denotes the function of PD.

We now discuss the details of UnX and FeE.

**A. THE UNCERTAINTY EXPLORATION MODULE**

The uncertainty exploration module UnX aims at finding more complete foreground regions from the uncertain areas while eliminating the interferences from the backgrounds according to the saliency distribution from the transformer encoder. The salient regions represent significant foreground objects and, therefore, their features can be taken as a guidance to divide the whole feature map into three types areas: foreground, background and uncertainty. Foreground represents the high-response areas, i.e., the salient foreground; background represents the low-response areas, i.e., the background objects apparently dissimilar to the foreground; and uncertainty is for features with the remaining, intermediate responses, i.e., representing the uncertain areas.

The uncertain areas can be both foreground and background objects and thus should be explored again for more salient features belonging to the foreground. Those features can be incorporated with the salient features from the backbone and consequently help boosting the discrimination of the network to recognize the foreground more completely and accurately.

Now the question is how to obtain the uncertain areas. Naively, a binary mask may be used to obtain them only, e.g., setting the low and high-response areas to zero. However, boosting the foreground features with such an equal setting to both types of areas may lead to some background areas to be fake detected again. Therefore, it’s better to have a mask than can help suppressing the low-responses areas. Consequently, a ternary guidance mask is defined, where the pixels corresponding to the low-response areas and high-response ones are set to -1 and 0 respectively with the pixels corresponding to the uncertain ones set to 1. This setup will make the network suppress the low-responses areas so that no background objects can be detected. It also emphasizes the uncertain areas by the highest mask values.

UnX is just designed based such a ternary mask (Fig. 4). The current feature map $U^c$ from previous low-resolution modules is firstly up-sampled and then normalized by the $sigmoid$ function to obtain a feature map $U^s$.

$$U^s = \delta(U(U^c)), \quad (4)$$

where $\delta$ and $U$ denote the $sigmoid$ function and up-sampling operation respectively.

Then, a ternary guidance mask $M$ is generated from $U^s$ according to two thresholds $\delta_h$ and $\delta_l$. The value of $i$-th pixel $M_i$ is set to be 1, 0 or -1 according to its corresponding feature value in $U^s$ being between these two thresholds, higher than $\delta_h$ or lower than $\delta_l$, representing area types of uncertainty, foreground or background, respectively.

$$M_i = \begin{cases} 1, & \text{if } U^s_i \leq \delta_l \\ 0, & \text{if } \delta_l \leq U^s_i \leq \delta_h \\ -1, & \text{if } U^s_i \geq \delta_h \end{cases} \quad (5)$$

Once $M$ is ready, it will multiply with the features from current stage, $F^c$, to obtain an weighted feature map $F^t$. This multiplication significantly highlights the uncertain areas for further foreground feature extraction.

$$F^t_i = \max(F^c_i, 0) \times M_i \quad (6)$$

According to Eq. (6), the responses from the background areas will be inverted in $F^t$ and consequently their possibilities of being foreground objects is greatly reduced. At the same time, the foreground areas will be erased. As a result, the features from the potential foreground areas show up.

Finally, $F^t$ is $3 \times 3$ convoluted for further feature integration and then combined with the up-sampled features from the previous modules $U^c$ through element-wise addition. This map is convoluted again by a $3 \times 3$ filter for integration so that a novel UnX feature $U^{c^e}$ is output for the next module.

$$U^{c^e} = fc(fc(F^t) + U(U^c)), \quad (7)$$
3 and 5. Features extracted from these branches are integrated by the $\text{concat}$ operation, and further convoluted by a $3 \times 3$ filter to obtain the final output $F^c$. The whole process can be formulated as follows:

$$F^c = f_c(C(B_1(F^c), B_2(F^c), B_5(F^c))),$$

(8)

where $C$ and $B_i (i = 1, 3, 5)$ represent the $\text{concat}$ operation and functions of the branches with a max atrous rate $i$ respectively.

We now discuss the loss design.

**C. LOSS FUNCTION**

There are totally four side supervisions gradually integrated to the four stages of the framework (Fig. 3). Therefore, the overall loss can be written as

$$\mathcal{L} = \sum_{m=2}^{5} \mathcal{L}_m,$$

(9)

where $\mathcal{L}_m (m = 2, 3, 4, 5)$ are the corresponding side-supervision losses between the UnX features $U^{(m)}$ and the ground truth.

For each side supervision, two losses are considered: the binary cross entropy loss (BCE) [50] and Intersection over Union (IoU) loss [51]. The former is the most widely used loss function based on pixel-level constraints, while the latter aims at optimizing the global structure rather than focusing...
on individual pixels. Consequently, the losses for the \( m \)-th side supervision can be formulated as

\[
\mathcal{L}_m = \mathcal{L}_{m}^{BCE} + \mathcal{L}_{m}^{IoU},
\]

(10)

where \( \mathcal{L}_{m}^{BCE} \) and \( \mathcal{L}_{m}^{IoU} \) are the BCE and IoU losses respectively:

\[
\mathcal{L}_{m}^{BCE} = - \sum_i p_i \log \hat{p}_i + (1 - p_i) \log (1 - \hat{p}_i),
\]

(11)

and

\[
\mathcal{L}_{m}^{IoU} = 1 - \frac{\sum_i p_i \hat{p}_i}{\sum_i (p_i + \hat{p}_i - p_i \hat{p}_i)},
\]

(12)

where \( p_i \) and \( \hat{p}_i \) represent ground truth and the output from the proposed framework respectively for \( i \)-th pixel in the \( m \)-th stage output.

The whole framework including the new modules, UnX and FeE, has been introduced till now and next the training process to fulfill the segmentation is summarized.

IV. TRAINING

A. TRAINING SETTINGS

The proposed model is implemented in PyTorch [52] with NVIDIA RTX 2080Ti for GPU acceleration. Adam [53] algorithm was used to optimize the model parameters. The learning rate is set to 1E-4, while the batch size for training is 16 with the maximum epoch setting to 100. Thresholds \( \delta_i \) and \( \delta_i \) are set to 0.7 and 0.3 respectively.

B. TRAINING ALGORITHM

The whole process is trained in an end-to-end way as shown in Algorithm 1. Note that the \( j \)-th UnX module in this algorithm represents the UnX module directly connecting to the \( j \)-th encoder stage as shown in Fig. 3.

Algorithm 1 Algorithm for the Polyp Segmentation of Colonoscopy Images by Exploring the Uncertain Areas

**Input:** Image \( I \), Ground truth \( G \).

**Output:** The segmentation prediction, \( P \).

1: Initialize the transformer encoder \( T \) with pre-trained parameters and randomly initialize the remaining parts.

2: **while** training is not convergent **do**
3: Generate the backbone feature maps \( F(i) (i = 1, 2, 3, 4) \leftarrow T(I, G) \).
4: Generate the initial saliency map \( U^{(5)} \) (Eq. (3)).
5: Compute the loss \( \mathcal{L}_5 \) with \( U^{(5)} \) (Eq. (10)).
6: \( \mathcal{L} = \mathcal{L}_5 \).
7: \( j = 5 \).

8: **while** \( j > 2 \) **do**
9: \( \text{For the } j \text{-th UnX module,} \)
\[ \text{a. normalize its saliency map } U^{(j)} \text{ (Eq. (4));} \]
\[ \text{b. generate its ternary guidance mask (Eq. (5));} \]
\[ \text{c. generate its weighted feature map (Eq. (6));} \]
\[ \text{d. generate its output feature map (Eq. (7)).} \]
10: Compute the loss \( \mathcal{L}_{j-1} \) (Eq. (10)).
11: \( \mathcal{L} = \mathcal{L} + \mathcal{L}_{j-1} \) (Eq. (9)).
12: \( j = j - 1 \).
13: **end while**
14: Optimize network by minimizing \( \mathcal{L} \).
15: **end while**
16: Compute \( P \) by sigmoid, \( i.e., P = \delta(U^{(2)}) \).

selected from Kvasir [57] and CVC-ClinicDB [55]. The sizes of all training images are set to 352 × 352. All five datasets except the training images from Kvasir and CVC-ClinicDB are taken as testing images.

B. EVALUATION METRICS

Several popular metrics are adopted to evaluate the performances, including the Dice Similarity Coefficient (DSC) [58], Intersection over Union (IoU) [59], Weighted F-measure \( F_w^{\beta} \) [60], and Mean Absolute Error (MAE) [61].

DSC and IoU are similarity measures at the regional level, focusing on the internal consistency of the segmented objects:

\[
\text{IoU} = \frac{|A \cap B|}{|A \cup B|},
\]

(13)

\[
\text{DSC} = \frac{2|A \cap B|}{|A| + |B|},
\]

(14)

where \( A \) and \( B \) represent pixel sets for the ground truths and their detection results, respectively.

This paper additionally takes mDSC and mIoU to separately represent the average values of DSCs and IoUs of all \( n \) test images, \( i.e., \)

\[
m\text{IoU} = \frac{1}{n} \sum_{i=1}^{n} \text{IoU}_i, \]

(15)

\[
m\text{DSC} = \frac{1}{n} \sum_{i=1}^{n} \text{DSC}_i, \]

(16)
where IoU and DSC represent IoU and DSC for i-th test image respectively.

\[ F^w_β = \frac{(1 + \beta^2)P^w \cdot R^w}{\beta^2 \cdot (P^w + R^w)}, \]

(17)

where: \( P^w \) and \( R^w \) represent precision and weighted recall [60], respectively; and 2) \( \beta \) is the coefficient and set to one here.

MAE is a pixel-by-pixel comparison index, denoting the average absolute error between the predicted value and the true value. The smaller the MAE, the better the model.

\[ \text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |\hat{p}_i - p_i|, \]

(18)

where \( \hat{p}_i \) and \( p_i \) represent the prediction and corresponding ground truth for i-th pixel of total n pixels.

C. RESULTS

Three types of experiments are taken to show the performances of the proposed method: Qualitative and quantitative experiments, and ablation study. In the experiments, several methods including the state-of-the-arts methods are adopted, including U-Net [16], SFA [42], PraNet [3], UACANet-S [4] SANet [33] and MSNet [8]. For fair comparison, the segmentation results are not thresholded with 0.5.

1) Qualitative experiments

These experiments are on the visual comparisons of the proposed method with the existing methods including the state-of-the-arts ones. Fig. 6 shows some segmentation results of different methods. The proposed method can effectively detect the polyps of different shapes without blurry boundaries (e.g., 1st row and 2nd row) and fakes from the similar tissues in backgrounds (e.g., 3rd row). It can be seen that our results are closest to ground truths for all images among all methods.

Fig. 7 show the comparison of the segmentation results on some challenging images. Those polyps are either rather small (e.g., 2nd row) or with irregular shapes (e.g., 1st row). The proposed method again demonstrates strong segmentation abilities. Especially, for ETIS (3rd row), only our method can successfully locate the polyp without the affections of other tissues in the background.

2) Quantitative experiments

Statistical comparisons with existing methods are taken to quantitatively evaluate the effectiveness of the proposed method. Considering training sets are selected from CVC-ClinicDB and Kvasir, the comparisons on learning ability are undertaken firstly. As shown in Table 1, our model is optimal on both datasets, with mDSC 1.8% and 0.5% higher than the state-of-the-arts method MSNet in CVC-ClinicDB and Kvasir, respectively.

Next the generalization abilities are experimented with three unseen datasets (CVC-ColonDB, ETIS and CVC-300) (Table 2). The results of the proposed method are still optimal. Specially, for ETIS, the most challenging among the five datasets, where most images contain too small polyps to be easily find, our method achieves 7.7% and 5.6% improvements in mDSC and mIoU respectively, comparing with the state-of-the-arts method MSNet.

Fig. 8 shows the DSCs of these methods under different thresholds on five polyp segmentation benchmark datasets. These curves show that the proposed model consistently outperforms other models, which proves its good capability for polyp segmentation.

3) Ablation study

The quality of the proposed method, especially the performances of UnX and FeE are evaluated through this study. Different configuration of the propose method are considered here, denoted with different notations as follows:

- 'Baseline' for the backbone network consisting of the transformer encoder and PD;
- 'Baseline+UnX+FeE' for the proposed full model;
- 'w/o UnX' and 'w/o FeE' representing the removal of UnX or FeE from 'Baseline+UnX+FeE' (the full model) respectively.

First, the feature abstraction abilities of different configurations are visually tested, where the feature maps are extracted from the last convolution before PD and uniformly re-sized to 44 x 44 for better display. As can be seen in Fig. 9, the proposed network (full model) can detect more complete polyps than the network without UnX (w/o UnX). This is because the UnX can mine more polyp regions in the uncertain areas efficiently. And compared to the network without FeE (w/o FeE), our proposed network (full model) can capture clearer details to enhance the feature representation.

Then the statistical comparisons of the performances under these configurations are taken (Table 3). As can be seen, except on Kvasir, the results from w/o UnX or w/o FeE are lower than those of the proposed full model. For Kvasir, the results from the full model are only slightly lower than those from w/o FeE. This is due to the fact that there are more other surrounding tissues on Kvasir dataset, so the network with FeE may extract those features which can interfere the feature responses of polyps.

VI. CONCLUSIONS AND DISCUSSIONS

There are two main challenges in polyp segmentation of colonoscopy images: blurry boundaries and close resemblances between polyps and surrounding tissues. To overcome these two difficulties, a new transformer encoder based polyp segmentation network is introduced, which takes a powerful ternary guidance mask based module, Uncertainty eXploration (UnX), to disclose more latent foreground areas from the uncertain areas and thus obtain robust foreground responses for complete object localization without
FIGURE 6. Visual comparisons of the polyp segmentations among different methods. (a) Origin images; (b) GT represents the ground truth; and (c) - (i) represent segmentation results from different methods.

FIGURE 7. More qualitative comparisons of the polyp segmentations for some challenging images. (a) Origin images; (b) GT represents the ground truth; and (c) - (i) represent segmentation results from different methods.

TABLE 1. Statistical comparisons of the learning abilities on CVC-ClinicDB and Kvasir. The best results are shown in bold.

| Method       | CVC-ClinicDB | Kvasir |
|--------------|--------------|--------|
|              | mDSC | mIoU | $F_β^n$ | MAE | mDSC | mIoU | $F_β^n$ | MAE |
| U-Net [16]   | 0.823 | 0.755 | 0.811   | 0.019 | 0.818 | 0.746 | 0.794   | 0.055 |
| SFA [42]     | 0.700 | 0.607 | 0.647   | 0.042 | 0.723 | 0.611 | 0.670   | 0.075 |
| PraNet [3]   | 0.899 | 0.849 | 0.896   | 0.009 | 0.898 | 0.840 | 0.885   | 0.030 |
| UACANet-S [4] | 0.907 | 0.855 | 0.902   | 0.009 | 0.897 | 0.840 | 0.884   | 0.028 |
| SANet [33]   | 0.916 | 0.859 | 0.909   | 0.012 | 0.904 | 0.847 | 0.892   | 0.028 |
| MSNet [8]    | 0.921 | 0.879 | 0.914   | 0.008 | 0.907 | 0.862 | 0.893   | 0.028 |
| Ours         | **0.939** | **0.891** | **0.940** | **0.007** | **0.912** | **0.859** | **0.908** | **0.027** |
TABLE 2. Statistical comparisons of the generalization abilities on CVC-ColonDB, ETIS and CVC-300. The best results are shown in bold.

| Method     | CVC-ColonDB | ETIS | CVC-300 |
|------------|-------------|------|---------|
|            | mDSC | mIoU | F\_\beta | MAE | mDSC | mIoU | F\_\beta | MAE | mDSC | mIoU | F\_\beta | MAE |
| U-Net [16] | 0.512 | 0.444 | 0.498 | 0.061 | 0.398 | 0.355 | 0.366 | 0.036 | 0.710 | 0.627 | 0.684 | 0.022 |
| SFA [42]   | 0.469 | 0.347 | 0.379 | 0.094 | 0.297 | 0.217 | 0.231 | 0.109 | 0.467 | 0.329 | 0.341 | 0.085 |
| PraNet [3] | 0.709 | 0.640 | 0.696 | 0.045 | 0.628 | 0.507 | 0.600 | 0.031 | 0.871 | 0.797 | 0.843 | 0.010 |
| UACANet-S [4] | 0.772 | 0.690 | 0.752 | 0.035 | 0.680 | 0.599 | 0.613 | 0.026 | 0.893 | 0.822 | 0.867 | 0.006 |
| SANEt [33] | 0.753 | 0.670 | 0.726 | 0.043 | 0.750 | 0.654 | 0.685 | 0.015 | 0.888 | 0.815 | 0.859 | 0.008 |
| MSNet [8]  | 0.755 | 0.678 | 0.737 | 0.041 | 0.719 | 0.664 | 0.678 | 0.020 | 0.869 | 0.807 | 0.849 | 0.010 |
| Ours       | 0.800 | 0.721 | 0.789 | 0.028 | 0.796 | 0.720 | 0.766 | 0.015 | 0.908 | 0.845 | 0.894 | 0.006 |

FIGURE 8. Comparison of the DSC curves under different thresholds. (a)-(e) represent the DSC curves on CVC-ClinicDB, ETIS, CVC-300, CVC-ColonDB and Kvasir respectively.

FIGURE 9. Feature extraction comparison of our method under different configurations. (a) Image and the corresponding ground truth; (b) features extracted by different configurations. It can be seen that removing any module can lead to incomplete features.
the distraction of tissues in the backgrounds. A multi-scale feature augmentation module, FeE, is also incorporated into the framework to obtain enhanced multi-scale features for UnX, so that varying sizes of shapes of polyps can be coped with efficiently. A series of quantitative and qualitative experiments show that the proposed method is superior than the state-of-the-arts methods and can achieve robust polyp segmentation of colonoscopy images without the affections of blur boundaries and surrounding tissues.

However, the inconsistent color distributions across different colonoscopy image datasets may affect the segmentation results. In addition, some images with an extremely small proportion of polyps, i.e., as small as less than 0.1 of the total image area, are also challenges for the segmentation performance. In the future, new tactics to overcome these limitations will be figured out, e.g., new data enhancement methods or the inclusion of the weighed loss to balance the contributions from both background and foreground pixels.

**REFERENCES**

[1] P. Favoriti, G. Carbone, M. Greco, F. Pirozzi, R. E. M. Pirozzi, and F. Corcione, “Worldwide burden of colorectal cancer: A review,” *Updates in Surgery*, vol. 68, no. 1, pp. 7–11, 2016.

[2] J. J. Granados-Romero, A. I. Valderama-Treviño, H. E. Contreras-Flores, B. Barrera-Mera, M. Herrera Enriquez, K. Uriarte-Ruíz, J. Céballos-Villalba, A. G. Estrada-Mata, C. Alvarado Rodríguez, and G. Arauz-Peña, “Colorectal cancer: A review,” *Journal of International Medical Research*, vol. 5, no. 11, pp. 4667–4676, 2017.

[3] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, “PlaNet: Parallel reverse attention network for polyp segmentation,” in *MICCAI*, 2020, pp. 263–273.

[4] T. Kim, H. Lee, and D. Kim, “UACNet: Uncertainty augmented context attention for polyp segmentation,” in *ACM Multimedia*, 2021, pp. 2167–2175.

[5] X. Guo, C. Yang, Y. Liu, and Y. Yuan, “Learn to threshold: ThresholdNet with confidence-guided manifold mixup for polyp segmentation,” *IEEE transactions on medical imaging*, vol. 40, no. 4, pp. 1134–1146, 2021.

[6] S. Alam, N. K. Tomar, A. Thakur, D. Jha, and A. Rauniar, “Automatic polyp segmentation using U-Net-ResNet50,” in *MediaEval Workshop*, 2020.

[7] N.-Q. Nguyen and S.-W. Lee, “Robust boundary segmentation in medical images using a consecutive deep encoder-decoder network,” *IEEE Access*, vol. 7, pp. 33 795–33 808, 2019.

[8] X. Zhao, L. Zhang, and H. Lu, “Automatic polyp segmentation via multi-scale subtraction network,” in *MICCAI*, 2021, pp. 120–130.

[9] N.-Q. Nguyen, D. M. Vo, and S.-W. Lee, “Contour-aware polyp segmentation in colonoscopy images using detailed upsampling encoder-decoder networks,” *IEEE Access*, vol. 8, pp. 99 495–99 508, 2020.

[10] R. Zhang, G. Li, Z. Li, S. Cui, D. Qian, and Y. Yu, “Adaptive context selection for polyp segmentation,” in *MICCAI*, 2020, pp. 253–262.

[11] N. C. Thanh, T. Q. Long et al., “CRF-EfficientUNet: An improved unet framework for polyp segmentation in colonoscopy images with combined asymmetric loss function and CRF-RNN layer,” *IEEE Access*, vol. 9, pp. 156 987–157 001, 2021.

[12] X. Guo, Z. Chen, J. Liu, and Y. Yuan, “Non-equivalent images and pixels: Confidence-aided resampling with meta-learning mixup for polyp segmentation,” *Medical Image Analysis*, vol. 78, p. 102394, 2022.

[13] M. Fiori, P. Musé, and G. Sapiro, “A complete system for candidate polyps detection in virtual colonoscopy,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 28, no. 07, p. 1460014, 2014.

[14] A. V. Mamonov, I. N. Figueiredo, P. N. Figueiredo, and Y.-H. R. Tsai, “Automated polyp detection in colon capsule endoscopy,” *IEEE Transactions on Medical Imaging*, vol. 33, no. 7, pp. 1488–1502, 2014.

[15] R. Nawarathna, J. Oh, J. Muthukadage, W. Tavanapong, J. Wong, P. C. De Groen, and S. J. Tang, “Abnormal image detection in endoscopy videos using a filter bank and local binary patterns,” *Neurocomputing*, vol. 144, pp. 70–91, 2014.

[16] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *MICCAI*, 2015, pp. 234–241.

[17] Z. Zhou, M. M. R. Siddiquiee, 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.

[18] M. Yeung, E. Sala, C.-B. Schönlieb, and L. Rundo, “Focus u-net: A novel dual attention-gated cnn for polyp segmentation during colonoscopy,” *Computers in Biology and Medicine*, vol. 137, p. 104815, 2021.

[19] J. Zhong, W. Wang, H. Wu, Z. Wen, and J. Qin, “PolypSeg: An efficient context-aware network for polyp segmentation from colonoscopy videos,” in *MICCAI*, 2020, pp. 285–294.

[20] Q. Xu, W. Duan, and N. He, “DCSAU-Net: A deeper and more compact Split-attention U-Net for medical image segmentation,” *arXiv preprint arXiv:2202.00692*, 2022.

[21] Y. Zhang, H. Liu, and Q. Hu, “TransFuse: Fusing transformers and CNNs for medical image segmentation,” in *MICCAI*, 2021, pp. 15–24.

[22] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, and L. Shao, “Polyp-PVT: Polyp segmentation with pyramid vision transformers,” *arXiv preprint arXiv:2108.06932*, 2021.

[23] S. Li, X. Sui, X. Luo, X. Xu, Y. Liu, and R. S. M. Goh, “Medical image segmentation using squeeze-and-expansion transformers,” in *IJCAI*, 2021.

[24] J. Wang, Q. Huang, F. Tang, J. Meng, J. Su, and S. Song, “Stepwise feature fusion: Local guides global,” *arXiv preprint arXiv:2203.05635*, 2022.

[25] Y. Lee, J. Kim, J. Willette, and S. J. Hwang, “Mpvit: Multi-path vision transformer for dense prediction,” in *CVPR*, 2022.

[26] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, “Pyramid Vision Transformer: A versatile backbone for dense prediction without convolutions,” in *ICCV*, 2021, pp. 568–578.

[27] W. Wang, W. Meng, and X. Fan, “Flexible and Accurate Salient Object Detection,” in *ECCV*, 2020, pp. 234–250.

[28] X. Sun, P. Zhang, D. Wang, Y. Cao, and B. Liu, “Coloroectal polyp...
segmentation by U-Net with dilation convolution,” in ICMLA, 2019, pp. 851–858.

[31] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, and H. D. Johansen, “ResUnet++: An advanced architecture for medical image segmentation,” in ISM, 2019, pp. 225–2255.

[32] F. E. Li, R. VanRullen, C. Koch, and P. Perona, “Rapid natural scene categorization in the near absence of attention,” Proceedings of the National Academy of Sciences, vol. 99, no. 14, pp. 9596–9601, 2002.

[33] J. Wei, Y. Hu, R. Zhang, Z. Li, S. K. Zhou, and S. Cui, “Shallow attention network for polyp segmentation,” in MICCAI, 2021, pp. 699–708.

[34] D. Wang, S. Chen, Q. Chen, Y. Cao, B. Liu, X. Liu, and X. Sun, “AFP-Mask: Anchor-free polyp instance segmentation in colonoscopy,” IEEE Journal of Biomedical and Health Informatics, 2022.

[35] Q. Hou, P.-T. Jiang, Y. Wei, and M.-M. Cheng, “Self-Embracing for integral object attention,” in NeURIPS, 2018, pp. 547–557.

[36] O. H. Maghsoudi, “Superpixel based segmentation and classification of polyps in wireless capsule endoscopy,” in SPMB, 2017, pp. 1–4.

[37] X. Jia, X. Mai, Y. Cui, Y. Yuan, X. Xing, H. Seo, L. Xing, and M. Q.-H. Meng, “Automatic polyp recognition in colonoscopy images using deep learning and two-stage pyramid feature prediction,” IEEE Transactions on Automation Science and Engineering, vol. 17, no. 3, pp. 1570–1584, 2020.

[38] D. Vázquez, J. Bernal, F. Sánchez, G. Fernández-Esparrach, A. López, A. Romero, M. Drozdal, and A. Courville, “A benchmark for endoluminal scene segmentation of colonoscopy images,” Journal of Healthcare Engineering, vol. 2017, pp. 4037190–4037190, 2017.

[39] N. Ibtehaz and M. S. Rahman, “Multiresunet: Rethinking the U-Net architecture for multimodal biomedical image segmentation,” Neural Networks, vol. 121, pp. 74–87, 2020.

[40] H. A. Qadir, Y. Shin, J. Solhusvik, J. Bergsland, L. Aabakken, and I. Balasingham, “Polyp detection and segmentation using mask R-CNN: Does a deeper feature extractor CNN always perform better?” in ISMIMICT, 2019, pp. 1–6.

[41] 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 EMBC, 2019, pp. 7223–7226.

[42] Y. Fang, C. Chen, Y. Yuan, and K.-y. Tong, “Selective feature aggregation network with area-boundary constraints for polyp segmentation,” in MICCAI, 2019, pp. 302–310.

[43] R. Wang, S. Chen, C. Ji, J. Fan, and Y. Li, “Boundary-aware context neural network for medical image segmentation,” Medical Image Analysis, vol. 78, p. 102395, 2022.

[44] K.-B. Park and J. Y. Lee, “SwinE-Net: Hybrid deep learning approach to novel polyp segmentation using convolutional neural network and swing transformer,” Journal of Computational Design and Engineering, vol. 9, no. 2, pp. 616–632, 2022.

[45] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” in ICLR, 2016.

[46] 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.

[47] D. Anguelov, D. Erhan, J. Szegedy, S. Eslami, and A. Toshev, “Beyond pixels: Weak supervision for image segmentation,” in CVPR, 2019, pp. 65–73.

[48] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in ICLR, 2021.

[49] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in CVPR, 2016, pp. 770–778.

[50] K. Simonyan and A. Zisserman, “Very deep convolutional networks for largescale image recognition,” in ICLR, 2015.

[51] K. P. Murphy, Machine learning: A probabilistic perspective. MIT press, 2012.

[52] M. A. Rahman and Y. Wang, “Optimizing Intersection-Over-Union in deep neural networks for image segmentation,” in ISVC, 2016, pp. 234–244.

[53] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in PyTorch,” in NeURIPS Workshop, 2017.

[54] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in ICLR, 2015.

[55] 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.

[56] D. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilatela, “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.

[57] 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.

[58] 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 MMM, 2020, pp. 451–462.

[59] N. J. Tutison and J. C. Gee, “Introducing Dice, Jaccard, and other label overlap measures to ITK,” Insight Journal, vol. 2, p. 707, 2009.

[60] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes challenge: A retrospective,” International Journal of Computer Vision, vol. 111, no. 1, pp. 98–136, 2015.

[61] R. Margolin, L. Zelnik-Manor, and A. Tal, “How to evaluate foreground maps?” in CVPR, 2014, pp. 248–255.

[62] T. Chai and R. R. Draxler, “Root mean square error (RMSE) or mean absolute error (MAE),” Geoscientific Model Development Discussions, vol. 7, no. 1, pp. 1525–1534, 2014.
ENMING ZHANG is currently working as an associate professor at the Department of Clinical Science Malmö in Lund University. He received the PhD degree in biomedical engineering in 2005 from Zhejiang University and then had his first postdoc training 2006-2007 at Nanyang Technological University in Singapore. From 2007-2010, he was granted a Marie-Curie postdoc fellowship to study L-type calcium channels in disease at the Department of Clinical Science, Faculty of Medicine at Lund University. After that, he obtained an assistant professor position (Forskassistent) to study ion channels and islet cell dysfunctions. From 2014, he received a permanent research position working at experimental endocrinology at the Department of Clinical Science Malmö in Lund University. His research group currently focuses on investigating pancreatic islet function and its relevance to type-2 diabetes using advanced technologies, e.g. nanotechnology, super-resolution microscopy (dSTORM) and super-bright beam resources such as synchrotron radiation.

***