Abstract

Colonoscopy is a common and practical method for detecting and treating polyps. Segmenting polyps from colonoscopy image is useful for diagnosis and surgery progress. Nevertheless, achieving excellent segmentation performance is still difficult because of polyp characteristics like shape, color, condition, and obvious non-distinction from the surrounding context. This work presents a new novel architecture namely Adaptation of Distinct Semantics for Uncertain Areas in Polyp Segmentation (ADSNet), which modifies misclassified details and recovers weak features having the ability to vanish and not be detected at the final stage. The architecture consists of a complementary trilateral decoder to produce an early global map. A continuous attention module modifies semantics of high-level features to analyze two separate semantics of the early global map. The suggested method is experienced on polyp benchmarks in learning ability and generalization ability, experimental results demonstrate the great correction and recovery ability leading to better segmentation performance compared to the other state of the art in the polyp image segmentation task. Especially, the proposed architecture could be experimented flexibly for other CNN-based encoders, Transformer-based encoders, and decoder backbones.

1 Introduction

Image segmentation is a major and significant topic in computer vision. This task classifies each pixel of the incoming image into predetermined classifications. Medical image segmentation is one of the highlight applications of segmentation techniques such as polyp segmentation, brain tumor segmentation, skin lesion segmentation, or lung segmentation.

Due to its complexity, polyp segmentation has recently attracted a lot of interest. Polyps are abnormal tissue growth from the surface of internal organs and can be found in the colon, rectum, stomach, or even throat. In most cases, polyps are benign, which means that they do not indicate illness or maliciousness. However, since polyps are capable of developing into cancer, a long-term diagnostic is necessary to determine whether or not they have become malignant. Therefore, identifying polyps in the colonoscopy image is helpful in facilitating the early detection of polyp-related diseases. Colonoscopy is the primary method
to locate and remove polyps, but this procedure requires a significant amount of time. In addition, there are still some challenges in a practical setting: common polyps can differ in size, color, and shape. Besides, polyps can develop haphazardly or densely in numerous sites, and may be challenging to distinguish them from surrounding tissues. This requires a reliable endoscopic polyps segmentation approach with high segmentation efficiency in difficult contexts.

Figure 1: Segmentation example of our model. Red boxes refer to uncertain areas. Purple boxes stand for noise details.

In general, difficult details and hazy areas pose restrictions to medical image segmentation architectures. Previous deep learning-based works focused on the result at the final stage without paying attention to modifying and recovering these areas. Therefore, modern architecture systems struggle to identify and categorize challenging features and unclear areas, as shown in Figure 1. The area indicated in the red box of other works is fairly dark and separates from the nearby polyp items, in this case, the model could not identify sufficient desired objects. Similarly, areas in the purple box are confused with actual polyps, causing misclassification. Under our supervision, these areas might be correctable and recoverable, since they gradually weaken and are diluted with decoder progress. Motivated by this, we propose a new approach, Adaptation of Distinct Semantics for Uncertain Areas in Polyp Segmentation (ADSNet), that uses a complementary trilateral decoder to give a superior early global map. Proposing a new module: continuous attention, we fix inaccurate details and re-store missed areas following two semantics: background semantic and object semantic. This strategy improves learning ability, generalization capability and overcomes the weaknesses of current models in challenging contexts.

The following sections are organized as follows: In section II, we present outstanding methods for salient object detection and associated reports to the polyp segmentation task. Section III explores the ADSNet components, while section IV executes experiments to observe the performance of the proposed solution. In the final section, we summarize the entire contributions of the paper.

2 Related Work

Salient Object Detection. SOD is a crucial preprocessing method for many computer vision applications including semantic segmentation, visual tracking, and image retrieval. Traditional SOD solutions are mainly based on heuristic priors (such as color, texture, and contrast) to construct saliency maps. With the advancement of Deep Neural Networks (DNNs), salient object detection (SOD) \cite{8} \cite{23} has made significant strides. First, several methods pay attention to improve the accuracy, Edge Guidance Network (EGNet) \cite{29} combines each path from the top-down stream for the salient object branch and the edge detection branch.
BANet [19] employed side-out fusion and single-stream, respectively, for boundary and object branches. To forecast salient objects with complete structure and exquisite borders, AFNet [10] proposed a multi-scale attentive feedback model and Boundary-Enhanced Loss. Other approaches consider striking a balance between accuracy and efficiency. RAS [5] explicitly multiplies a reverse prediction region to acquire the remaining features for saliency refinement. CPD [25] introduced a cascade partial decoder that uses an attention strategy to enhance high-level features while discarding shallow layer features for acceleration. ITS-DNet [31] proposed an interactive two-stream decoder to investigate several cues, such as saliency, contour, and their interaction.

Polyp Segmentation. With the development of deep learning techniques, the polyp segmentation task is explored strongly. Brandao et al. [4] proposed a fully convolutional neural network to gain great polyp segmentation performance. The first U-shape architecture Unet [17] with two paths: an encoder to extract features and a decoder to educate the final map proved excellent performance in biomedical image segmentation. Motivated by the U-shape structure, several advanced versions were introduced. ResUnet++ [11] utilized the advantages of residual link, channel attention, and Atrous Spatial Pyramidal Pooling. Unet++ [32] kept rich information by dense concatenation via multi-stages. Several other variants use multi U-shape like DoubleUnet [12], and CUNet [21] to achieve better feature representation. Newer methods focus on the relationship between area and boundary, SFANet [9] considered area and border constraints. PraNet [8] applied Reverse Attention Module to correct the boundary area to boost the segmentation performance. Inspired by PraNet [8], UACANet [14] was designed with a new attention: parallel axial attention and augmenting uncertain area to model border information. SSFormer [22] used a pyramid Transformer encoder to improve the generalization ability of models and propose a new decoder to emphasize local features. Another transformer-based architecture, Polyp-PVT [7] proposed a similarity aggregation module to extract local pixels and global semantic cues from the polyp area, effectively suppressing noise in features and significantly improving their expressive capabilities.

3 Methodology

The overall proposed architecture is visualized in Figure 2. The network extracts four levels of spatial features \( \{ f_i : i = 1, \ldots, 4 \} \). The complementary trilateral decoder [30] constructs an early global map from encoder features. The continuous attention produces two distinct semantic masks from weak and strong regions before concatenating with the early global map to give a superior final mask. Each component will be explained in further detail in the following subsections.

3.1 Encoder Backbone

In computer vision tasks in general, the encoder component extracts features that bring important semantics for analyzing objects. Recent works often used CNN-based, Transformer-based encoders or combined both in several particular situations. CNN-based approaches do well in extracting local features by using the local kernel, Transformer-based architectures are effective in grasping global relationships, improving generalization ability, and multi-scale feature processing ability, but they need more data or a strong pre-trained. In a differ-
quent approach without transformer, our work uses a CNN-based encoder to extract features and exploit the features obtained by analyzing separate semantics. This also brings many advances in generalization ability and multi-scale feature processing ability. In particular, we use EfficientNet-V2S [20] as an encoder backbone that produces multiple levels of spatial features $f_i \in \mathbb{R}^{H_{i+1} \times W_{i+1} \times C_i}$ where $C_i \in \{256, 512, 1024, 2048\}$ and $i \in \{1, 2, 3, 4\}$.

3.2 Complementary Trilateral Decoder

Deep features offer important information for locating objects, while shallow features have high resolution and do not store many valuable textures. Therefore, we focus on utilizing high-level features with no regard for low-level features. In this work, we introduce a Complementary Trilateral Decoder [30] - a new SOTA decoder for silent object detection to address the loss of spatial structure, lack of boundary detail, and diluted semantic context. The early global map is obtained by $M = CTD(f_1, f_2, f_3, f_4)$. By estimating and using a decision parameter, we divide the early global map into two weak and strong regions to analyze separated semantics. The first component is strong (S) areas that hold clear object structure, and the second one is weak (W) areas that are determined by the decision parameter and refer to uncertain areas that could be reconstructed.

![Figure 2: The proposed ADSNet architecture.](image)

3.3 Continuous Attention for Background Semantic and Object Semantic

The multi-scale representation can assist in perceiving multi-scale objects for SOD. Particularly, polyps frequently come in a variety of sizes, hence we suggest a unique structure called Progress Atrous Spatial Pyramidal Pooling (PASPP) [27] to progressively capture multi-scale high-level features $\{f_2, f_3, f_4\}$. Lower-level features $\{f_1\}$ often hold rich detail
information in appearance, we adapt Camouflage Identification Module (CIM) [7] to represent better texture and edge information. The CIM [7] is operated by the following two consecutive attention mechanisms:

\[
\begin{align*}
\text{Att}_c &= \sigma (H_1 (G_{\text{max}}(x)) + H_2 (G_{\text{avg}}(x))) \otimes x \\
\text{Att}_s &= \sigma (\text{Conv}3x3(\text{Cat}(R_m(x), R_a(x)))) \otimes x
\end{align*}
\]

Where \(x\) is the input feature. \(G_{\text{max}}, G_{\text{avg}}\) are global max pooling and global average pooling functions, respectively. \(H_1, H_2\) are two convolutional layers to reduce the dimension and recover the original dimension. \(R_m, R_a\) are max pooling and average pooling following channel dimension. \(\text{Cat}\) is concatenate operation, while \(\sigma\) stands for Sigmoid function.

Background Semantic. At the early stage, the beginning mask rough prediction objects without clear structural details. To boost the segmentation performance, the details around the objects need to be detected and classified more precisely. Several previous works used reverse attention[5] to highlight the outside area of objects to process the boundary constraint without correcting inaccurate details through texture information at low-level features. While low-level features often hold rich detail information, such as texture, color, edges, and polyps frequently resemble the background in appearance. Inspired by that motivation, we utilize \(f_1\) to guide high-level features, modify inaccurate details around polyp objects and highlight strong components (S) to give a background semantic mask (BS)

\[
BS = \text{Conv}3x3(\text{Conv}3x3(\text{Cat}(S \otimes AG_i), i = 2, 3, 4))
\]

Object Semantic. Recent architectures cannot segment sufficiently recognized objects by creating a final mask only with an encoder and a decoder. Following our findings, sev-
eral areas the model can detect, but they are so weak and easy to vanish through decoder progress because of the highlight of strong details, or detected areas but very noisy with the surrounding space. In addition, polyp colonoscopy images are one of the most difficult objects to segment because of the similarity between objects and backgrounds. Discovering weak features or challenging objects will have the potential to give outstanding performance. Therefore, instead of only using the strong (S) component as in the background semantic, we combine weak (W) and strong (S) components to explore and recover uncertain areas.

\[ OS = \text{Conv}3x3(\text{Conv}3x3(\text{Cat}((S,W) \otimes AG_i), i = 2, 3, 4)) \] (4)

To implement the two ideas proposed above, we introduce Continuous Attention as depicted in Figure 3 including gate attention mechanisms that are executed consecutively to modify the semantics of input high-level features \(\{f_2, f_3, f_4\}\) before concatenating them to provide OS and BS as described in Figure 2.

### 3.4 Loss Function

When analyzing polyp images, the consideration of boundary and area contributes greatly to segmentation performance, so we propose using a loss function described below. Whereas ACE loss \([\Box]\) offers advantages in terms of geometrical limitations, compact curvature, length, and area of active contours with region similarity, whereas BCE loss \([\Box]\) represents a pixel restriction. We define the loss function as follows:

\[\text{Loss}(y, \bar{y}) = \text{ACE}(y, \bar{y}) + \text{BCE}(y, \bar{y})\] (5)

Where \(\bar{y}, y\) denote the predicted mask and ground truth, respectively.

**ACE Loss**

\[\text{ACE}(y, \bar{y}) = (\alpha + \beta \bar{K}^2)|\nabla \bar{y}| + \lambda y(c_1 - \bar{y})^2 + \lambda (1 - y)(c_2 - \bar{y})^2\] (6)

Where \(\alpha\) and \(\beta\) execute the trade-off of length and curvature. \(\lambda\) is used to balance two clauses. \(\bar{K}\) is the curvature of \(y\), \(c_1\) and \(c_2\) are mean intensities of the interior and outer regions.

**BCE Loss**

\[\text{BCE}(y, \bar{y}) = \sum_{i=1}^{n} y_i \log(\bar{y}_i) + (1 - y_i) \log(1 - \bar{y}_i)\] (7)

### 4 Experiments

This section describes the dataset, evaluation metrics, experimental results, and further insights into the output. From experimental results, we evaluate and compare ADSNet with current cutting-edge techniques for the polyp segmentation challenge.

**Dataset**

- Kvasir-Seg \([\Box]\) There are 1000 polyp images and related annotations. The sizes range from 332x487 to 1920x1072, and the polyps that appear in the images likewise have a range in size and shape.
• CVC-ClinicDB consists 612 images with the size of 384 × 288 from 25 colonoscopy videos.

• ETIS contains 196 images with the size of 1225 × 966 collected from 34 colonoscopy videos. Because polyps are frequently small and challenging to locate, this dataset is more challenging.

• CVC-ColonDB chooses 380 images from 15 different colonoscopy sequences.

Evaluation Metrics We use some standard metrics for the purpose of segmenting medical images. The Dice Similarity Coefficient (DSC) statistic and Intersection over Union (IoU) term are used to quantify the degree of similarity between two samples. MAE shows the average absolute error between the true mask and the anticipated mask. Weighted F-measure \( F_{\beta} \) measures the effect of false negative, false positive, and true positive.

4.1 Results

We set up the same training and testing with other methods which are compared in the next section: 1450 (900 images from Kvasir-Seg and 550 images from CVC-ClinicDB) for training. ETIS, CVC-ColonDB for generalization ability testing, and remaining images in Kvasir-Seg and CVC-ClinicDB for learning ability evaluation.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Kvasir-Seg</th>
<th>CVC-ClinicDB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dice</td>
<td>IoU</td>
</tr>
<tr>
<td>DCRNet</td>
<td>0.886</td>
<td>0.825</td>
</tr>
<tr>
<td>PraNet</td>
<td>0.898</td>
<td>0.840</td>
</tr>
<tr>
<td>SANet</td>
<td>0.904</td>
<td>0.847</td>
</tr>
<tr>
<td>Polyp-PVT</td>
<td>0.917</td>
<td>0.864</td>
</tr>
<tr>
<td>Ours</td>
<td>0.920</td>
<td>0.871</td>
</tr>
</tbody>
</table>

Table 1: Quantitative evaluation of diverse models on Kvasir-Seg & CVC-ClinicDB Datasets. The best results are bolded.

Learning ability. In this experiment, the domain of the test and train set is similar. We compare ADSNet with recent SOTA methods including: PraNet, DCRNet, SANet, and Polyp-PVT. For a fair comparison, all results of the models mentioned above are referenced from the original papers. The detailed results shown in Table 1 indicate that ADSNet improves recent approaches. 0.920 Dice score and 0.871 IoU score on Kvasir-Seg and 0.938, 0.890 on CVC-ClinicDB, respectively. Besides, our model also is better on \( F_{\beta} \) and MAE indexes. For qualitative results, Figure 4 shows obtained segmentation performance by all models on Kvasir-Seg and CVC-ClinicDB. The segmentation performance of competitors is referenced completely from public sources in Polyp-PVT. The great agreement between predicted samples by ADSNet and ground truths is illustrated. This proves the efficient and accurate segmentation ability of the proposed architecture in difficult and challenging situations with uncertain areas that previous approaches have not dealt with yet.
Generalization ability. The compared results are shown in Table 2. On ETIS [1] dataset, the performance obtain 0.798 Dice score and 0.715 IoU score improved by 1.39% and 1.27% over Polyp-PVT [7] respectively. ADSNet also shows great generalizability in CVC-ColonDB [18] when surpassing all other SOTA methods with the highest scores, 0.815 Dice and 0.730 IoU score. Although on CVC-ColonDB [18] the $F_w^\beta$ reduces when compared with SANet [24], our proposal still improves on the remaining. Several noteworthy and difficult samples are shown in Figure 5. It is clear that ADSNet is capable of partitioning complex and minute details in obscure situations in addition to segmenting well in challenging cases with uncertain areas.
<table>
<thead>
<tr>
<th>Train Set</th>
<th>Kvasir-Seg &amp; CVC-ClinicDB</th>
<th>Test Set</th>
<th>CVC-ColonDB</th>
<th>ETIS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Methods</td>
<td>Dice</td>
<td>IoU</td>
<td>$F_w^\beta$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DCRNet</td>
<td>0.704</td>
<td>0.631</td>
<td>0.684</td>
<td>0.052</td>
</tr>
<tr>
<td>PraNet</td>
<td>0.712</td>
<td>0.640</td>
<td>0.699</td>
<td>0.043</td>
</tr>
<tr>
<td>SANet</td>
<td>0.753</td>
<td>0.670</td>
<td><strong>0.892</strong></td>
<td>0.043</td>
</tr>
<tr>
<td>Polyp-PVT</td>
<td>0.808</td>
<td>0.727</td>
<td>0.795</td>
<td>0.031</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>0.815</strong></td>
<td><strong>0.730</strong></td>
<td>0.860</td>
<td><strong>0.029</strong></td>
</tr>
</tbody>
</table>

Table 2: Quantitative evaluation of diverse models on unseen Datasets ETIS & CVC-ColonDB. The best results are bolded.

4.2 Further insights

This part proves the effectiveness of the combination of OS, BS, and M in the final mask. When the background semantic (BS) modifies wrong details around and highlights the primary structure of objects. The object semantic (OS) explores ambiguous and challenging areas. We display several segmentation performances in Figure 6. It can be seen that OS and BS assist the model capture context information in both object and background semantics. Meanwhile, without OS or BS, the architecture has difficulty in detecting and categorizing uncertain details, causing misclassification as in the last column. We also visualize the output feature map of the early, background semantic, and object semantic in Figure 7 to verify the effectiveness of analyzing two distinct semantics, ADSNet offers significantly more sufficient features.
5 Conclusion

In this paper, we have proposed a new novel architecture called ADSNet for exploring uncertain areas to improve polyp segmentation performance. We have introduced a new SOTA decoder that utilizes high-level features from the CNN-based encoder to produce an early global map. Finally, we have analyzed two separate semantics of the early global map: background semantic and object semantic using continuous attention to give an excellent map. The obtained results demonstrate the superiority of the proposed model in dealing with challenging cases with uncertain areas, from that, obtain better Dice, IoU, $f_w$, and MAE scores over recent state-of-the-art methods.

Acknowledgements This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00256629) grant funded by the Korea government(MSIT). This work was also supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF2021R111A3A04036408). This work was supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub). The corresponding author is Soo-Hyung Kim.

References


