scientific reports

Check for updates

OPEN

# NA-segformer: A multi-level transformer model based on neighborhood attention for colonoscopic polyp segmentation

Dong Liu[1,2,3,5], Chao Lu[1,2,5], Haonan Sun[1,4] & Shouping Gao[1,2]✉

**In various countries worldwide, the incidence of colon cancer-related deaths has been on the rise in recent years. Early detection of symptoms and identification of intestinal polyps are crucial for improving the cure rate of colon cancer patients. Automated computer-aided diagnosis (CAD) has emerged as a solution to the low efficiency of traditional methods relying on manual diagnosis by physicians. Deep learning is the latest direction of CAD development and has shown promise for colonoscopic polyp segmentation. In this paper, we present a multi-level encoder-decoder architecture for polyp segmentation based on the Transformer architecture, termed NA-SegFormer. To improve the performance of existing Transformer-based segmentation algorithms for edge segmentation on colon polyps, we propose a patch merging module with a neighbor attention mechanism based on overlap patch merging. Since colon tract polyps vary greatly in size and different datasets have different sample sizes, we used a unified focal loss to solve the problem of category imbalance in colon tract polyp data. To assess the effectiveness of our proposed method, we utilized video capsule endoscopy and typical colonoscopy polyp datasets, as well as a dataset containing surgical equipment. On the datasets Kvasir-SEG, Kvasir-Instrument and KvasirCapsule-SEG, the Dice score of our proposed model reached 94.30%, 94.59% and 82.73%, with an accuracy of 98.26%, 99.02% and 81.84% respectively. The proposed method achieved inference speed with an Frame-per-second (FPS) of 125.01. The results demonstrated that our suggested model effectively segmented polyps better than several well-known and latest models. In addition, the proposed method has advantages in trade-off between inference speed and accuracy, and it will be of great significance to real-time colonoscopic polyp segmentation. The code is available at https://github.com/promisedong/NAFormer.**

**Keywords** Deep Learning, Segmentation, Computer Vision, Colonoscopy, Colon Cancer, Transformer, Image Processing

Colorectal cancer (CRC) is one of the most common malignant tumors. According to the latest cancer statistics worldwide, colorectal cancer currently ranks third in both incidence and mortality[1]. CRC usually begins as non-cancerous protrusions (called polyps) within the tissue lining the colon. CRC has a high good survival rate in the early stages, but once it reaches the later stages, survival rates drop dramatically. Therefore, early diagnosis and effective treatment are essential to reduce the occurrence of colon cancer. Colonoscopy is considered the gold standard for detecting colorectal lesions, enabling early detection and removal of colorectal polyps to prevent disease progression[2,3]. Currently, colonoscopy are mainly performed manually by professional radiologists, who are primarily limited by their experience, efficiency, and the complexity of the polyps itself. Therefore, automated polyp segmentation technology aims to precisely identify polyps in the early stages, which is of great significance for improving diagnostic efficiency and preventing colon cancer.

Generally, the color of colonic polyps is similar to that of surrounding tissue, and the presence of intestinal mucus can further obscure their appearance, resulting in vague boundaries[4]. Successful segmentation of colonic polyps, particularly small-target polyps, is, therefore, a challenging task. Recently, with the help of computer

[1]Hunan Engineering Research Center of Advanced Embedded Computing and Intelligent Medical Systems, Xiangnan University, Chenzhou 423300, China. [2]School of Computer and Artificial Intelligence, Xiangnan University, Chenzhou 423300, China. [3]Key Laboratory of Medical Imaging and Artificial Intelligence of Hunan Province, Xiangnan University, Chenzhou 423300, China. [4]College of Software, Jilin University, Changchun 130012, China. [5]Dong Liu and Chao Lu contributed equally to this work. ✉email: gaoshoup@xnu.edu.cn

vision technology, many deep learning networks have been proposed and applied to colonic polyp segmentation with remarkable results. Current deep learning segmentation solutions can be divided into two types: convolutional neural network (CNN)-based and Transformer-based. The first type of methods mainly combine some competitive architecture or modules in the framework of convolutional neural networks to improve learning ability. For example, the U-shaped architecture a remarkable "encoder–decoder" architecture[5]has gained significant attention for its ability to leverage multi-level features to reconstruct high-resolution prediction outcomes. R. Zhang et al[6]. designed a convolutional neural network (CNN) that employed an attention mechanism, global context module, and adaptive selection module to adaptively process different context information, considering the diverse shapes and sizes of colonic polyps along with the low contrast between the polyps and the surrounding tissue. Recently, another deep learning framework, Transformer[7]has gained great attention and achieved significant success in medical image segmentation. Some representative Transformer variants, such as Swin-Unet[8], Trans-Unet[9]and SETR[10], tried to add transformer layers or optimize the structure to capture global dependencies and shown great potential in polyp segmentation. A recent representative method based on Transformer framework in image segmentation was Missformer[11], which designed a novel feed-forward network and ReMixed Transformer Context Bridge that achieved remarkable experimental results.

Although deep learning methods show great image processing capabilities, achieving accurate colonic polyp segmentation is a difficult yet worthy endeavor[12]. First, some colonic poly data are not adequately annotated, and the available dataset samples are small that affect the effectiveness of deep learning models. Second, due to the small difference between polyp tissue and surrounding healthy tissue, accurate identification of polyp boundaries is challenging. Furthermore, although robust global modeling capabilities of Transformers can enhance the segmentation performance of pixel-level segmentation, its lack of local context information in modeling results in limited performance in polyp segmentation applications. The standard multi-headed attention mechanism requires the input image to be sliced into image patches, leading to a lack of continuity between patches and an increase in computational complexity. Especially, based on the characteristics of colon polyp images, the neighbor context is helpful to improve the results of polyp segmentation, but it has not been fully explored.

Therefore, this study is devoted to develop a Transformer encoder-decoder network based on neighbor attention mechanism and unified focal loss to tackle the issue of poor edge segmentation performance in polyp segmentation. More specifically, we proposed a novel and effective segmentation method NA-SegFormer, which is a cutting-edge Transformer-based semantic segmentation framework that prioritizes effectiveness, precision, and robustness. With the neighbor attention mechanism embed in our model, more local context and global–local correlations can be captured. On the other hand, unified focal loss was introduced to effectively address the issue of category imbalance in polyp datasets. To summarize, the key innovations of our methodology are as follows:

1. NA-SegFormer, a multilevel segmentation approach with neighborhood feature embedding and rare category feature exploiting, was proposed based on Transformer architecture.
2. To enhance model performance, we introduced NAPM, a new module that utilized the neighbor attention mechanism and overlapping patch fusion to improve edge segmentation accuracy in the Transformer framework.
3. We performed extensive evaluations on three challenging benchmark datasets and our model achieved best segmentation performance.

The remainder of this article is structured as follows. First, the related work is introduced in Section "Related work". The proposed NA-SegFormer model for polyp segmentation is described in Section "Method". The experimental results are analyzed in Section "Experiments and Analysis", and we conclude in Section "Conclusions".

## Related work

### Semantic segmentation

Long et al[13].[14] were among the first to propose a deep learning method for semantic image segmentation, using a fully convolutional network (FCN). The FCN model showed a 20% improvement over conventional methods on the Pascal VOC 2012 dataset. To address the limitation of the FCN model, which neglected global context information, Liu et al[15]. introduced the ParseNet model. ParseNet enhances FCNs with global context by incorporating the average feature of a layer to enhance features at each location. Despite the popularity and productivity of the standard FCN model, it has several drawbacks, including slow real-time reasoning, limited consideration of global background information, and difficulty in conversion to three-dimensional (3D) images. Chen et al[16]. proposed a semantic segmentation approach that utilized fully connected conditional random fields (CRFs) and CNNs. They highlighted that the final layer of deep CNNs does not provide sufficiently localized responses for accurate object segmentation. To overcome this, they combined the outputs of the final CNN layer with a fully connected CRF. Similarly, Badrinarayanan et al[17]. introduced SegNet, a promising convolutional encoder-decoder architecture for image segmentation. Notably, SegNet's decoder employed non-linear upsampling using pooling indices obtained during the associated encoder's max-pooling step to upsample lower-resolution input feature maps. HRNet[18]has gained considerable attention given its ability to maintain high-resolution representations throughout the encoding process by linking high-to-low-resolution convolution streams parallelly and transferring information regularly between resolutions. HRNet is the backbone of many current semantic segmentation models, which employ contextual models such as self-attention and its extensions to enhance performance. Given the extensive research on attention mechanisms in computer vision, it is unsurprising that studies have applied these mechanisms to semantic segmentation. For instance, Fu et al[19]. proposed a dual attention network based on the self-attention mechanism for scene segmentation to capture rich contextual relationships. The Transformer is a deep neural network that is based on self-attention

and was originally developed for natural language processing[20,21]. In the past two years, the Transformer structure and its modifications have been successfully applied to computer vision[22]. Zheng et al[23]. were the first to use the Transformer for semantic segmentation and developed the Segmentation Transformer Network (SETR) to extract global semantic information. Due to the requirement of serialization of the Transformer input and output, SETR separates two-dimensional images into patches and transforms them into one-dimensional sequences. Position encoding is applied to retain spatial information while learning the specific coding of each patch. Strudel et al[7]. developed Segmenter, a pure Transformer model for semantic segmentation tasks. Each encoder layer in the Segmenter captures global context information. The mask transformer then decodes the encoder output and performs class embedding to produce a three-dimensional segmentation feature map, which is subsequently upsampled to generate the final segmentation map.

## Medical image segmentation

Medical image segmentation plays a crucial role in comprehending medical images, extracting, and recognizing features, and providing a quantitative assessment of lesions or anomalies. Accurate segmentation of skin lesions is essential for the identification of melanoma and subsequent cancer diagnosis and planning of the therapeutic course. Wang et al[24]. proposed an innovative Boundary-Aware Transformer (BAT) that utilizes prior knowledge about lesion borders to design a boundary-aware attention gate in the Transformer architecture. This attention gate provides feedback for efficient training of BAT with supplementary boundary-wise supervision. Dental radiography is commonly used for detecting disorders such as periapical lesions and bone abnormalities. Lee, J.H., et al[25]. utilized CNN to identify dental caries in premolars, molars, or both using 846 dental notes with 30 panoramic radiographs for training and 20 radiographs for validation and testing. A region-based CNN (R-CNN) recognizes and locates dental features accurately[26]. Accurate segmentation of kidney tumors by computer-aided diagnosis systems is crucial for reducing the workload of radiologists and performing surgical procedures associated with these tumors. Shen et al[27]. proposed a hybrid encoder-decoder architecture, COTR-Net, comprising convolution and transformer layers for end-to-end segmentation of kidneys, kidney cysts, and kidney tumors. Despite great advances in medical image segmentation, polyp segmentation remain challenging because colonoscopic polyps vary in shape and have low contrast with surrounding tissue[28–30]. Therefore, to achieve efficient colonoscopic polyp segmentation requires further optimization of deep learning models combined with domain knowledge.

## Polyp segmentation

Early detection and treatment of colonoscopic polyps is an effective way to avoid colon cancer[31,32]. Recently, many polyp segmentation or detection methods have been developed to assist physicians during the colonoscopy. Based on a comprehensive literature survey, the current methods of colonoscopic polyp segmentation can be divided into three groups of visual techniques.

The first type of polyp segmentation method is based on convolutional neural network architecture. For example, Akbari et al[33]. proposed a polyp segmentation model based on a fully CNN that achieved superior results compared to conventional solutions. Qadir et al[34]. evaluated the polyp segmentation performance of different modern convolutional neural networks in the MASK R-CNN framework. Recently, Unet[5]employed an "encoder-decoder" architecture based on fully CNN and showed great potential in polyp segmentation. Since then, numerous variants of Unet have been developed to improve the performance of polyp segmentation. Jha et al[35]. introduced the ResUNet + + architecture and advocated for a DoubleUNet framework to perform the segmentation task. The DoubleUNet consists of two stacked U-Nets, and variants of this architecture have been frequently employed for polyp segmentation[36,37]. Mahmud et al[38]. proposed a novel encoder-decoder based modified deep neural network architecture, which utilized multi-scale contextual information and achieved promising performance in polyp segmentation. The second colonoscopic polyp segmentation method is based on the Transformer architecture. Chen et al[39]. proposed a hybrid network that cascades CNN and Transformer to extract both local and global features. The decoder upsamples the feature map and the skip connections achieve precise positioning. Xiao et al[40]. proposed a novel contrastive Transformer network (CTNet) for polyp segmentation, involved with self-multiscale interaction and collection information, which shows a convincing learning ability. Krenzer et al[41]. developed a transformer network-based polyp detection and classification system to enable assisted diagnosis. The latest representative network, namely Att-PVT[42], that combines CNN and Pyramid Vision Transformer (PVT) together for poly segmentation. In this work, multilevel feature fusion module and cascaded context integration strategy were employed to improve edge segmentation. In addition, some methods focus on real-time polyp segmentation, that is, optimizing lightweight network and improving segmentation speed. Jha et al[43]. proposed a novel architecture, namely NanoNet, which performed on a pre-trained MobileNetV2 combined with a modified residual block to achieve real-time performance. Lin et al[44]. developed a novel lightweight transformer, that achieved rapid polyp segmentation performance in endoscopic images by using a substitution strategy in the encoder and incorporating a inter-block feature fusion module in the decoder. An open-source real-time polyp-detection system[45] was provided, in which different deep convolutional neural networks were evaluated for clinical application in colonoscopy.

Although previous studies have improved and applied to colonic polyp segmentation, and the performance has been significantly enhanced. However, accurate segmentation for colonoscopic polyp remains problematic. There is still plenty of room for improvement, owing to the subtle changes of polyp appearance, the complexity of pathological features, and the limitations of the number of samples. Therefore, to address the above problems, this study is devoted to propose a multi-level Transformer model based on the neighbor attention mechanism to enhance the edge segmentation of colonic polyps.

## Method

### Overall architecture

As shown in Fig. 1, our proposed method NASegFormer mainly consists of two components, an encoder-decoder CNN and a long sequence Transformer feature extractor.

In the encoding stage, the input image was firstly transformed to obtained sequence feature through Neighborhood Feature Embedding. And then the feature was encoded by the TransFormer Block. Finally, the probability distribution of the feature was calculated by Feed Forward Network (FFN). After repeating the process 3 times, the feature extraction of the output feature map of colonic polyps at different levels can solve the problem that it was difficult to segment the polyp tissue similar to normal tissue under colonoscopy. Besides, the Neighborhood Attention module in the TransFormer network was designed to enhance the semantic understanding of the model for the normal tissue and edge features. Meanwhile, CNN was used to extract multi-scale features in the encoding stage. In the decoding stage, the TransFormer structure performed forward feature mapping of the output long sequence features through FFN. Finally, the deep semantic features output by the encoder was converted into classification probability feature images through upsampling and convolution operations.

**1) TransFormer Block with FFN:** It was composed of LayerNorm, Self-Attention and FFN. The role of Self-Attention was to perform normalized weight mapping of the original features to achieve a better feature matching degree and improve the feature extraction ability of the model. The TransFormer Block with FFN is formulated as Eq. (1)-(5):

$$Attention(Q, K, V) = LayerNorm(SoftMax(\frac{Q \times K^T}{\sqrt{d_{head}}}) \times V) \tag{1}$$

$$FFN(x_{in}) = GELU(LayerNorm(FC(x_{in}))) \tag{2}$$

$$Transformer\ Block(x_{in}) = Attention(x_1, x_2, x_3) + x_{in} \tag{3}$$

$$Output(x_{in}) = Transformer\ Block(x_{in}) \tag{4}$$

$$Transformer\ Block\ with\ FFN(x_{in}) = Output(x_{in}) + FFN(x_{in}) \tag{5}$$

where, $x_{in} \in \Re^{H \times W \times C}$ represents the input feature map, $Q, K, V$ is the feature sequence obtained by dividing $x_{in}$ according to the feature channel dimension and the shape size is $N \times C$, where $N = H \times W$, and $d_{head}$ is the number of channels.

**2) TransFormer Feature Encoding:** The TransFormer Encoder was composed of Neighborhood Feature Embedding and Transformer Block with FFN. For an input feature map $x_{in} \in \Re^{H \times W \times C}$, it would be divided into patches firstly through a convolution kernel with kernel size $7 \times 7$ and stride 4 to maintain the continuity of local sematic information of the polyp image. The Neighborhood Attention Module of the same branch, with the same kernel size and stride, was used to extract the neighboring global features of the patches. And the output

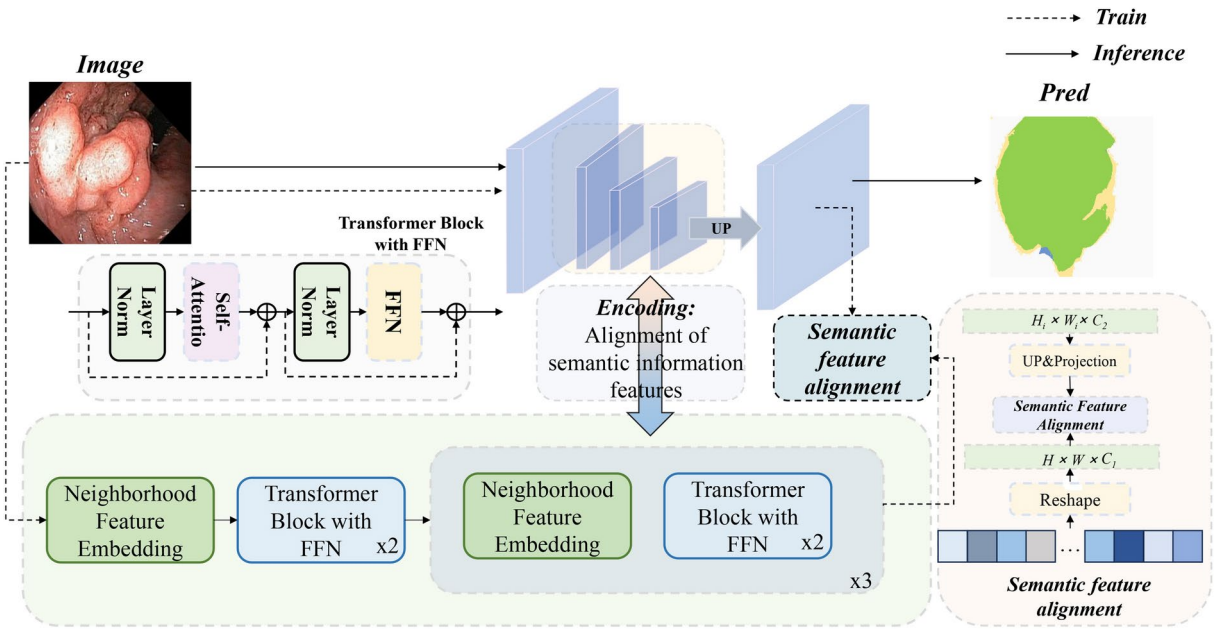


**Fig. 1**. The overall frame diagram of the NASegFormer network (Green is the Transformer architecture diagram, where the white dashed box corresponds to the TransFormer block with FFN, the orange box refers to the semantic distribution alignment of the decoding stage).

feature map $\vec{f} \in \Re^{H\times W\times C1}$ was then input to the next 3 transformer feature encoder, of which the kernel size, stride and padding were all $2 \times 2$, 3 and 1 respectively. This process can be formulated as Eq. 6.

$$Feature_{output} = \frac{H_{in}}{2^{i+1}} \times \frac{W_{in}}{2^{i+1}} \times 2^{i-1}C, (i = 1, 2, 3, 4) \tag{6}$$

where, $H_{in}, W_{in}$ represents the resolution of the feature map, and $C$ is for channel numbers. In the decoding stage, after a step-by-stage linear feature mapping, the feature map is upsampled to the original image scale, and finally convolution is used for category feature mapping. It can be formulated as Eq. 7.

$$Feature_{class}(x_{in}) = Conv_{class}(Up(Linear(x_{in}, C_{out}).transpose(B, C_{out}, \frac{H}{2^{i+1}}, \frac{W}{2^{i+1}}))), (i = 1, 2, 3, 4) \tag{7}$$

where, $x_{in}$ represents the input feature map, $class$ represents the number of input classes and $Conv$ is for convolution operation.

**3) Encoder-Decoder (CNN):** The network ResUnet was used as feature extractor. The output of the encoding stage was $Feat_i, i = (1, 2, 3, 4)$, which refers to Eq. 8. And the decoding stage was a stage-by-stage upsampling process until the size of the feature map same to the original input. At last, CNN was used to complete the classification regression, which can be defined as Eq. 9.

$$Feat_i = \frac{H_{in}}{2^i} \times \frac{W_{in}}{2^i} \times C_i, (i = 1, 2, 3, 4) \tag{8}$$

$$Sam_{class}(x_1, x_2) = Conv_{class}(cat(Up(x_1), x_2)) \tag{9}$$

where, $class$ represents the number of output classes, $x_1, x_2$ represents the shallow detail features and deep semantic features of the input respectively, and $cat$ is for patching operation of channels.

**4) Training Strategies:** In order to better express the TransFormer's ability of capturing long-distance contextual semantic information of image features, we first embed the Neighborhood Attention module in the TransFormer architecture to make up for the shortcomings of TransFormer's inability to obtain features from neighboring regions. Then, the output features were used as prior features to guide the CNN network, giving full play to the model's performance of local detail feature extraction, adjacent boundary discrimination and spatial context semantic understanding. In the training phase, mean square error was used to align the sequence features and semantic features for semantic expression.

## Neighborhood attention module

In colonoscopy polyp segmentation, the attention mechanism can enhance the effective interaction between local features and global features to a certain extent. Many colonoscopic polyp segmentation techniques employ attention mechanisms[42,46]. However, the previous attention mechanism has certain difficulties in segmenting normal polyp tissues and diseased tissues between adjacent regions, especially for the long sequence network of the TranFormer architecture, which destroys the continuity of local features. In this study, a neighborhood attention mechanism was designed obtain the global feature semantic information of neighboring regions by adding multi-scale pooling operation in the TransFormer encoding stage. As shown in Fig. 2, it can be specifically described as the high-frequency feature expression of the current neighboring features obtained by Max pooling of the input features respectively, so as to extract the shallow semantic information and deep semantic information $g_1 \in \Re^{H\times W\times C}$ of the boundary of normal polyp tissue and abnormal tissue. And the average pooling was used for extracting low-frequency feature $g_2 \in \Re^{H\times W\times C}$ of neighbor regions. After that, contextual semantic information of neighboring region was extracted through adaptivepool. The features $f_1 \in \Re^{H\times W\times C}$, $f_2 \in \Re^{H\times W\times C}$ of channel H and W was then added pixel by pixel and fine-grained feature extraction for redundant semantic features is performed to obtain $f_3 \in \Re^{H\times W\times C}$. Then the global feature $\widetilde{f} \in \Re^{H\times W\times C}$ that can express the features of neighboring regions was obtained through adding $g_1 \in \Re^{H\times W\times C}$ ,$g_2 \in \Re^{H\times W\times C}$,$f_3 \in \Re^{H\times W\times C}$ in channel dimension. After that, the convolution kernel with the same scale as the pooling was used for local feature extraction to make up for the feature response of the pooling to the local region. Meanwhile, in the other branch, the input features were transformed by Overlap Patch Embedding to output feature $T \in \Re^{H\times W\times C}$. Finally, feature $\widetilde{f}$ and $T$ were added pixel-by-pixel to obtain feature $P \in \Re^{H\times W\times C}$ , which combined local feature information and global information of neighboring regions.

## Unified focal loss for polyp segmentation

There are certain differences in the proportion of colonoscopy polyp size in the whole image, specifically, the foreground pixels in the small polyp tissue image account for too little of the whole image. Due to the imbalance of class distribution, the model lacks effective supervision for the foreground polyp tissue in the feature extraction, and is easy to favor the majority of categories for optimization in the training process. This will lead to negative optimization of the model and significantly affect the segmentation accuracy of the model. And in polyp medical images, the characteristics of high density and continuous characteristics are not obvious, especially in the high-frequency and low-frequency connection area. The traditional Focal loss can effectively suppress the background categories, but the Focal loss acts on all categories to reduce the loss contribution of small sample classes. Unified Focal loss selectively enhances or suppresses a certain type of features, overcomes the disadvantages of negative optimization of small sample classes and class gradient bias of background classes,

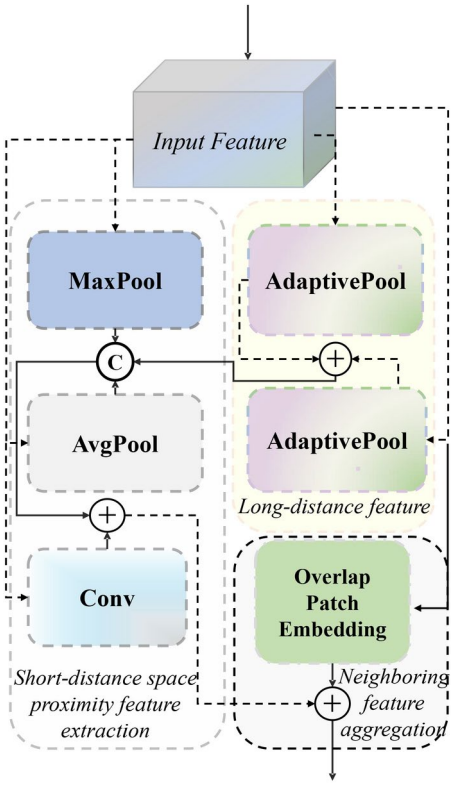


**Fig. 2**. Neighborhood Attention Module (NA).

and retains the suppression of background loss. In this paper, unified focal loss[47] wad used as segmentation loss function for segmentation supervision. It can be formulated as Eq. (10)-(17).

$$L_T = \sum_{b,i,j}^{B,H,W} \frac{(1-\widehat{y}_{b,i,j}+\varepsilon)}{(y_{b,i,j}\cdot\widehat{y}_{b,i,j}+\alpha\cdot y_{b,i,j}(1-\widehat{y}_{b,i,j})+(1-\alpha)\cdot(1-y_{b,i,j})\cdot\widehat{y}_{b,i,j}+\varepsilon)} \tag{10}$$

$$L_{back} = \sum_{c=0}^{C}(1-L_{T_c}) \tag{11}$$

$$L_{fore} = \sum_{c=1}^{C}(1-L_{T_c})\cdot(1-L_{T_c})^{\beta} \tag{12}$$

$$L_{ftl} = \frac{1}{N}\sum_{i,j}^{H,W}\sum_{c=1}^{C}(L_{back}+L_{fore}) \tag{13}$$

$$L_{ce} = \sum_{i=0,j=0}^{H,W}\sum_{c=0}^{C}\frac{y_{i,j,c}}{\log(\widehat{y}_{i,j,c})} \tag{14}$$

$$L_{ceb} = \sum_{i,j}^{H,W}\sum_{c=0}(1-\alpha)\cdot(1-\widehat{y}_{i,j,c})^{\beta}\cdot L_{ce}^{c} \tag{15}$$

$$L_{cef} = \sum_{i,j}^{H,W}\sum_{c=1}(\alpha\cdot L_{ce}^{c}) \tag{16}$$

$$L_{fl} = \frac{1}{N}\sum_{i=0,j=0}^{H,W}\sum_{c=0}^{C}(L_{ceb}+L_{cef}) \tag{17}$$

Finally our segmentation optimization strategy is to optimize the $L_{ftl}$ and $L_{fl}$ loss costs, as shown in Eq. 18.

$$L_{ufa} = \underset{\vartheta=0}{argmin}(L_{ftl} + L_{fl}) \tag{18}$$

## Experiments and analysis

### Experimental datasets and pre-processing

In this study, three challenging publicly available datasets were employed for the experimental evaluation, as shown below.

1. Kvasir-SEG[50]. It consists of a collection of 1,000 polyp images for polyps segmentation, which was prepared in Vestre Viken Health Trust in Norway.
2. Kvasir-Instrument[52]. It includes 590 images of endoscopic instruments with associated ground truth masks.
3. KvasirCapsule-SEG[43]. It consists of an open-access dataset that contains 55 images and corresponding ground truth masks for polyps segmentation, which is the polyp class of Kvasir-Capsule[51].

In order to present the experimental data more clearly, the image size, data augmentation strategy, training and testing data distribution and other relevant information were shown in Table 1. It is worth noting that we used an established division of training and test sets for Kvasir-Instrument, to keep the setup the same as the previous methods. Furthermore, we used Kvasir-Capsule as an unseen dataset to test the performance of the proposed method. Figure 3 depicts examples of polyps and their corresponding masks from the three datasets.

### Experimental setup

The hardware environment for this study consists of a computer configured with Intel Xeon Gold 6348 CPU and Nvidia A30 and Nvidia RTX 3090 PCIE GPUS. The software environment was developed on the Windows system and includes Python 3.7 and Pytorch1.13.1 + cu117 deep learning libraries. The details of the hyperparameters utilized are presented below.

For the experimental parameters, the batch size and training period were set to 64 and 32, respectively, and Adam[48] optimizer was employed for updating network parameters. We used cosine attenuation as the learning rate, the minimum learning rate is 1e-6, and the maximum learning rate is 2.5e-4. And the weight decay was set to 1e-3.

For the training strategy of Kvasir_instrument dataset and Kvasir_SEG dataset, a maximum of 35 K iterations were employed to ensure the each compared methods achieved best performance in the corresponding dataset. Furthermore, for SwinUnet and TransUnet, swin_tiny_patch4_window7_224.pth and imagenet21k + imagenet2012_R50 + ViT-B_16.npz were employed as pre-training weights, respectively. For our method, we use the same training strategy for Kvasir_instrument and Kvasir_SEG dataset. It is particularly worth noting that our model was trained from scratch, i.e., our model training was without pre-training, to ensure that the comparison results were fair enough. Since the sample number of the KvasirCapsule dataset was very small and not suitable for performing individual evaluation, we used it as unseen data to test the robustness of the model trained on the Kvasir-SEG dataset.

### Evaluation metrics

For the experimental evaluation, we utilize Dice Coefficient[49], Intersection over Union (IoU)[14], Precision, Recall, and Accuracy as the main evaluation indicators.

Dice Coefficient: also known as the F1 score, is the harmonic mean of precision and recall, and is defined as:

$$\text{Dice} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{19}$$

where TP (True Positive) is the number of correctly segmented polyp pixels, FP (False Positive) is the number of non-polyp pixels incorrectly labeled as polyp, and FN (False Negative) is the number of polyp pixels missed by the segmentation algorithm.

Intersection over Union (IoU): also known as the Jaccard Index, is the ratio of the intersection of the predicted and ground truth polyp regions to their union, and is defined as:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{20}$$

Precision: measures the percentage of correctly segmented polyp pixels among all the pixels that the algorithm predicted as polyp, and is defined as:

| | | Kvasir-SEG[50] | Kvasir-Instrument[52] | Kvasir-Capsule[43] |
|---|---|---|---|---|
| Data Processing strategy | Image resizing | 256 × 256 | 256 × 256 | 256 × 256 |
| | Data augmentation | horizontal flip, vertical flip,random rotation | horizontal flip, vertical flip, random rotation | — |
| Training strategy | Training | Random 80% | Established 80% | — |
| | Testing | Random 20% | Established 20% | All |

**Table 1**. Data processing strategy and Training data distribution for experimental evaluation.

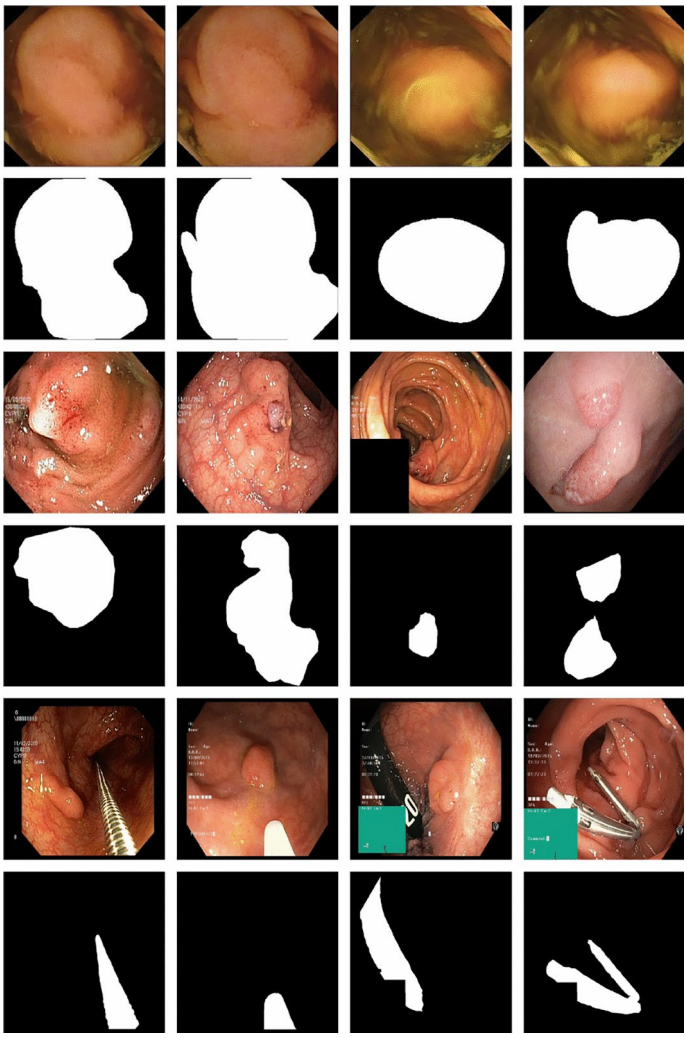

**Fig. 3**. Partial examples of the original images and the corresponding masks. The first two rows are examples from the KvasirCapsule-SEG dataset. The third and fourth rows are samples from the Kvasir-SEG dataset. The fifth and sixth rows are samples from the Kvasir-Instrument.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{21}$$

Recall: measures the percentage of correctly segmented polyp pixels among all the ground truth polyp pixels, and is defined as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{22}$$

Accuracy: measures the overall percentage of correctly segmented pixels, and is defined as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{23}$$

where TN (True Negative) is the number of correctly segmented non-polyp pixels.

## Comparison with the latest methods

In order to make a comprehensive comparison, eight popular approaches were selected for comparison, including four CNN-based methods (i.e., Unet[5], Unet++[53], ResUnet[35], and ResUnet++[35]) and four Transformer-based (i.e., SETR[10], SwinUnet[8], TransUnet[9], and MissFormer[11]). Half of these methods were published within the last three years, and all achieved the state of the art in image segmentation while publishing.

We constructed a series of experiments to evaluate our method against the above representative methods. First, We conducted a comparative evaluation on Kvasir-SEG and Kvasir-Instrument dataset, and KvasirCapsule-

| Model | Precision | Recall | Dice | mIoU | Accuracy |
|---|---|---|---|---|---|
| UNet | 83.71 | 82.12 | 82.91 | 82.45 | 94.84 |
| UNet++ | 60.86 | 91.79* | 73.19 | 75.19 | 93.33 |
| ResUNet | 86.75 | 91.66 | 89.14* | 88.38* | 96.84* |
| ResUNet++ | 72.69 | 90.40 | 80.59 | 80.80 | 94.76 |
| SETR | 56.87 | 81.41 | 66.96 | 70.58 | 91.61 |
| SwinUnet | 70.00 | 70.37 | 70.19 | 72.06 | 91.10 |
| TransUnet | 87.18* | 86.40 | 86.78 | 86.04 | 96.03 |
| MissFormer | 71.03 | 70.48 | 70.76 | 72.46 | 91.22 |
| NASegFormer (Ours) | **96.14** | **92.53** | **94.30** | **93.59** | **98.26** |

**Table 2**. Quantitative analysis on the Kvasir-SEG. Bolded parts represent better performance, and the parts marked with * indicate suboptimal.

| Model | Precision | Recall | Dice | mIoU | Accuracy |
|---|---|---|---|---|---|
| UNet | 91.74 | 91.39 | 91.56 | 91.36 | 98.43 |
| UNet++ | 89.13 | 91.59 | 90.34 | 90.23 | 98.23 |
| ResUNet | 86.91 | **99.01** | 92.57 | 92.38 | 98.71 |
| ResUNet++ | 88.61 | 92.60 | 90.56 | 90.44 | 98.29 |
| SETR | 88.75 | 91.64 | 90.17 | 90.07 | 98.21 |
| SwinUnet | 90.38 | 94.24 | 92.27 | 92.05 | 98.59 |
| TransUnet | 92.44* | 95.35 | 94.35* | 94.09* | 98.97* |
| MissFormer | 88.40 | 89.34 | 88.87 | 88.86 | 97.95 |
| NASegFormer (Ours) | **92.73** | 96.53* | **94.59** | **94.33** | **99.02** |

**Table 3**. Quantitative analysis for the Kvasir-Instrument. Bolded parts represent better performance, and the parts marked with * indicate suboptimal.

| Model | Precision | Recall | Dice | mIoU | Accuracy |
|---|---|---|---|---|---|
| UNet | 37.20 | 96.46 | 53.69 | 43.19 | 61.05 |
| UNet++ | 28.87 | 95.70 | 44.36 | 37.60 | 56.04 |
| ResUNet | 51.53 | 97.34 | 67.39 | 53.38 | 69.72 |
| ResUNet++ | 16.00 | 97.28 | 27.48 | 29.57 | 48.73 |
| SETR | 30.95 | 93.94 | 45.57 | 38.62 | 56.88 |
| SwinUnet | 33.36 | **98.03** | 49.78 | 40.95 | 59.14 |
| TransUnet | **85.51** | 96.36 | **90.61** | **80.23** | **89.24** |
| TransUnet# | 65.77 | 92.82 | 76.99 | 61.43 | 76.14 |
| MissFormer | 64.68 | 89.84 | 75.21 | 58.82 | 74.11 |
| NASegFormer(Ours) | 71.64* | 97.87* | 82.73* | 69.20* | 81.84* |

**Table 4**. Cross-validation qualitative analysis(training on Kvasir-SEG data, model generalization capability verification on KvasirCapsule-SEG data set). **Bolded** parts represent better performance, and the parts marked with * indicate suboptimal. # refers to TransUnet without using a pre-trained model.

SEG, and the results of the quantitative analysis are presented in Table 2 and Table 3, respectively. Then, to further compare the generalization ability and robustness of the models, KvasirCapsule-SEG was used as untrained data to test all comparison methods, and the cross-validation qualitative analysis results were shown in Table 4. To demonstrate the effectiveness of the proposed approach more intuitively, segmentation results of the compared methods for the Kvasir-SEG dataset, Kvasir-Instrument and KvasirCapsule-SEG are shown in Fig. 4, Fig. 5 and Fig. 6, respectively. In addition, we provide a visualization of the segmentation results coincident with GT for the compared methods, as shown in Fig. 7.

The following findings were obtained from the above experimental results.

Firstly, for the Kvasir-SEG dataset, the proposed method NASegFormer achieved the highest score for all evaluation metrics compared with other models; and for the Kvasir-Instrument dataset, our NASegFormer achieved the best performance on Precision, Dice, mIoU, and Accuracy, and the second best performance on

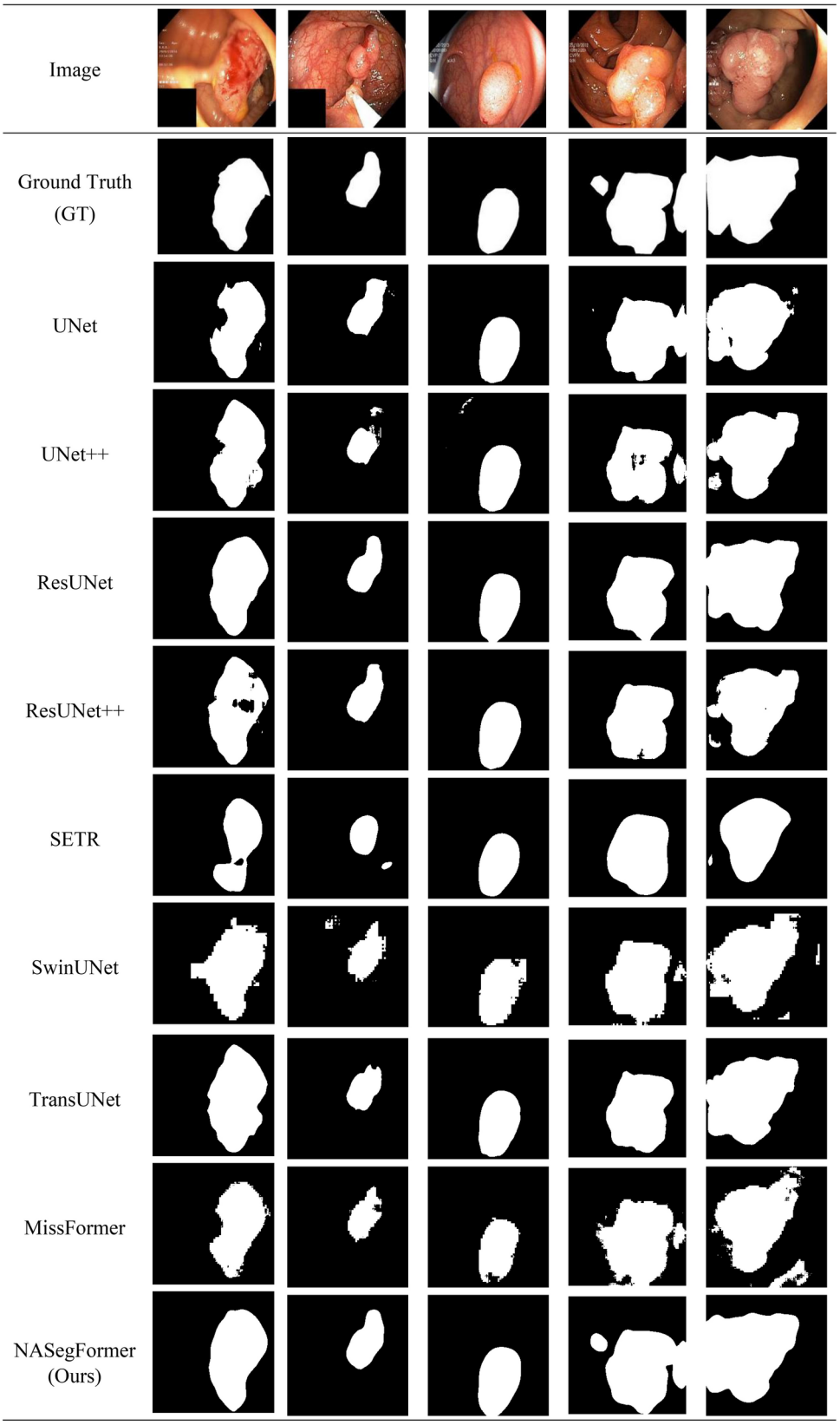


**Fig. 4**. Segmentation examples of the compared methods on the Kvasir-SEG dataset.

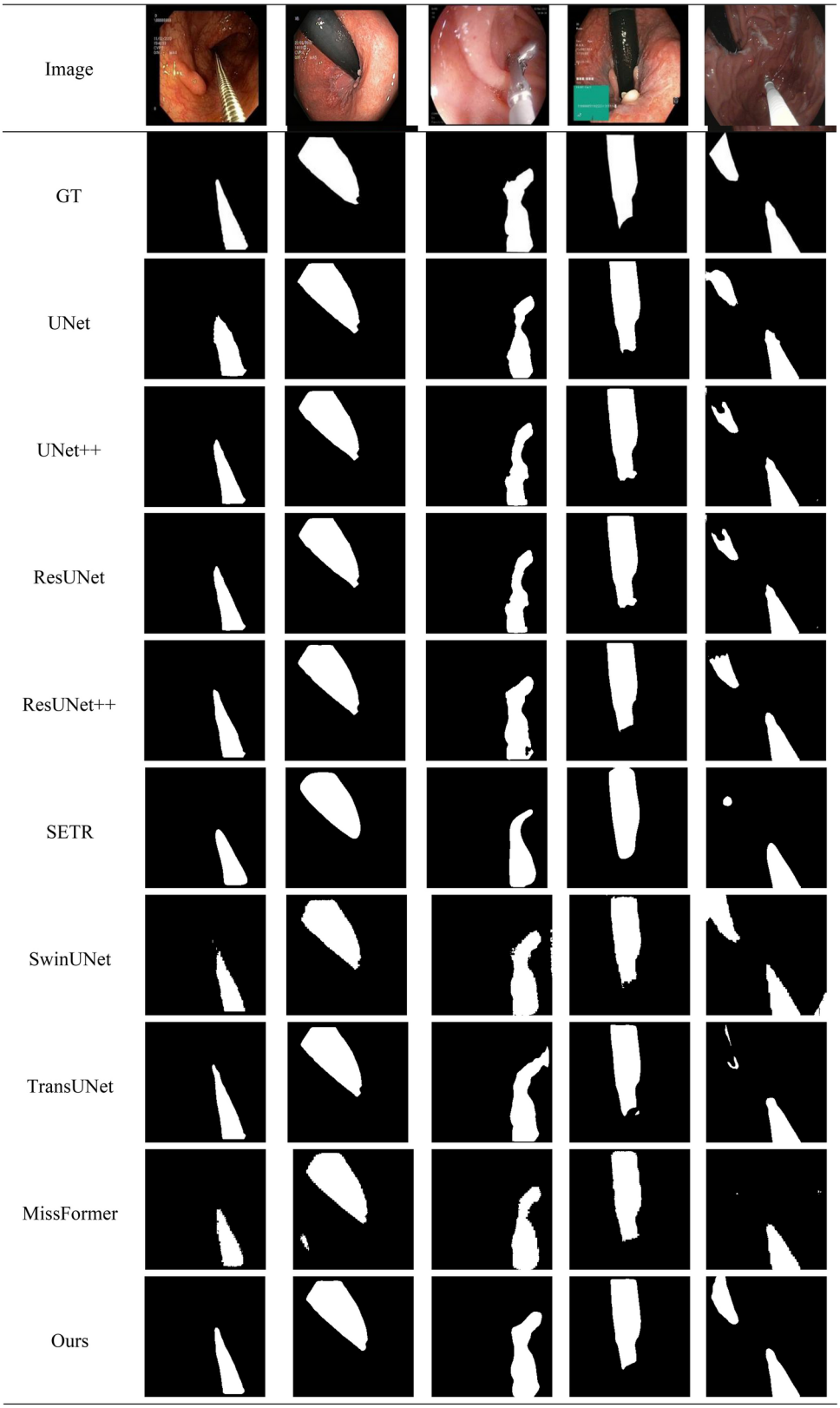


**Fig. 5**. Segmentation examples of the compared methods on the Kvasir-Instrument dataset.

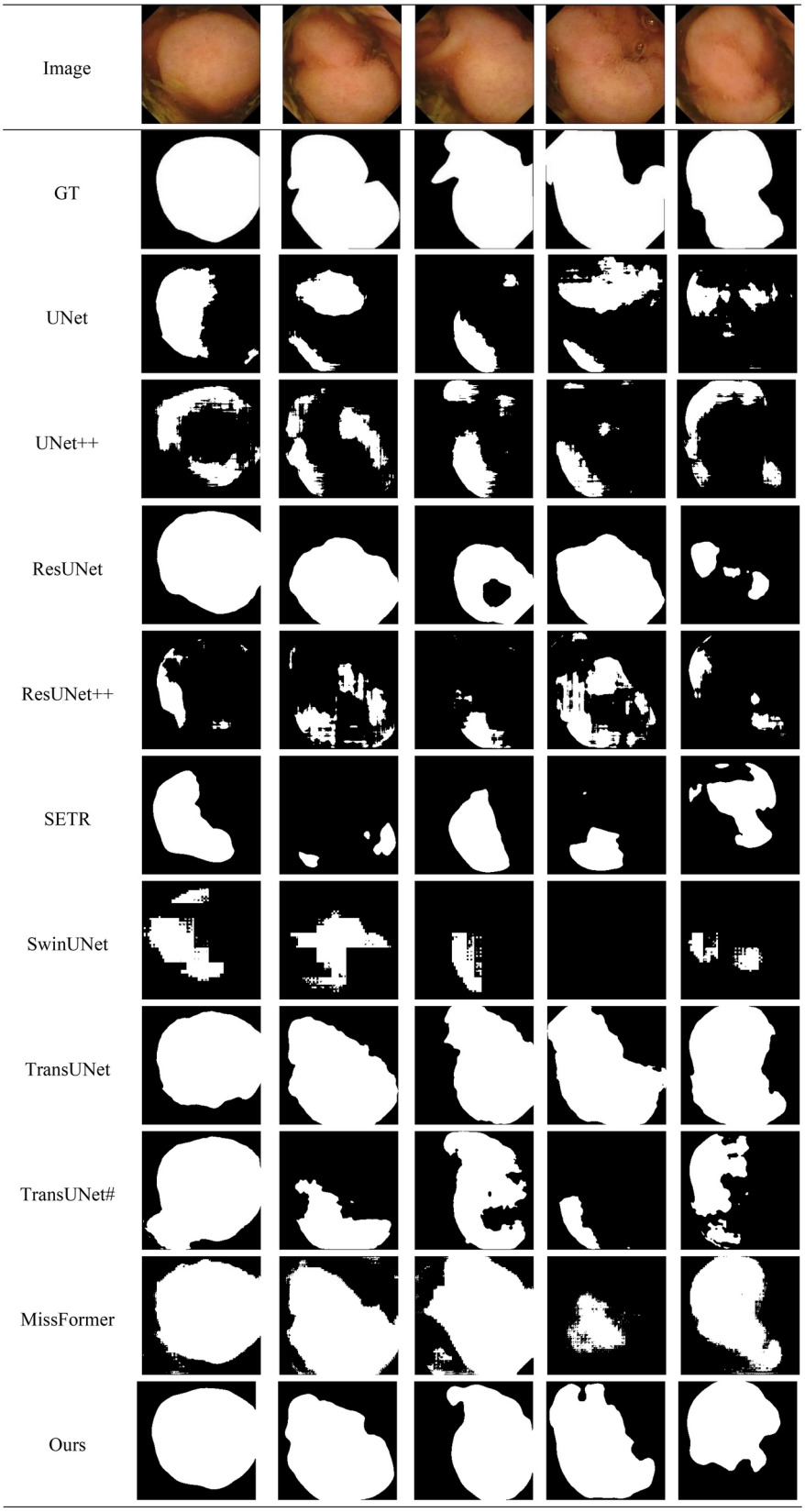


**Fig. 6**. Segmentation examples of the compared methods on the KvasirCapsule-SEG dataset.


**Fig. 7**. Visualization of the segmentation results coincident with GT for the compared methods on the Kvasir-SEG dataset (Green denotes TP; Blue denotes FP; Yellow denotes FN; Gray denotes TN).

Recall compared with other models. These results show the significant advantages of our method over other SOTAs (state-of-the-arts). More specifically, for the Kvasir-SEG dataset, our method outperformed ResUNet and TransUnet by 5.16% and 7.52%, respectively, with a Dice score of 94.30%. For the Kvasir-Instrument dataset, our proposed method achieved an accuracy of 99.02% with a Dice score of 94.59%, surpassing ResUNet by 2%.

In terms of comparison methods, ResUNet (CNN-based method) and TransUnet (Transformer-based method) achieved relatively high performance on the Kvasir-SEG and Kvasir-Instrument dataset, but there is still a certain gap from our method. This was mainly due to the fact that we improved the backbone networks with multiple optimization strategies. Especially, with the neighbor attention mechanism and unified focal loss, more discriminative features from both global and local perspectives could be captured to obtain a more accurate polyp segmentation results.

Secondly, by observing the results in Table 4 i.e. the cross-validation qualitative analysis results by training on Kvasir-SEG and testing on KvasirCapsule-SEG, our proposed model demonstrated Recall of 97.87%, Accuracy of 82.73, and Dice of 82.73%. In general, the performance of all compared methods in cross-validation experiments had decreased, which was due to the small number of samples in our training dataset Kvasir-SEG. Furthermore, the test data KvasirCapsule-SEG and the training data Kvasir-SEG were collected from different medical devices and different hospitals, and were labeled by different experts, that also presents a huge challenge for accurate segmentation of unseen data. Nevertheless, our approach has yielded competitive results. For example, we obtained sub-optimal performance in all indicators such as Precision, Recall, Dice, mIoU and Accuracy, which indicates that our model has good generalization ability and robustness. Another objective phenomenon that requires a greater explanation is that TransUnet achieved highest score in four evaluation metrics, such as Precision, Dice, mIoU and Accuracy, which are much higher than other methods. There are two possible reasons for this. On the one hand, new model like this published in recent years has good generalization ability; on the other hand, pre-trained model performed on a very large dataset ImageNet was used during the training process of TransUnet, which has a great impact on the improvement of segmentation performance. It should be emphasized that our approach still achieved sub-optimal results without using a pre-trained model. If we do not use the pre-trained model for the TransUnet method, but use the same criteria as ours, i.e., TransUnet$^{\#}$ in Table 4, our experimental results will exceed TransUnet$^{\#}$ across the board.

Thirdly, by observing the visual comparison on the Kvasir-SEG dataset in Fig. 4, various segmentation errors such as redundances, speckles, adhesions and holes exist in the segmentation results for different compared methods. However, the segmentation results of the proposed NASegFormer were very close to Ground Truth (GT). A more intuitive comparison was shown in the Fig. 4, which shows the degree of regional overlap between the segmentation results of all methods and GTs, and it can be found that our method achieves the best segmentation effect. Another strong evidence can be seen in the fourth column of the sample in Fig. 4, which has a single tiny polyp tissue in the GT, only our method completely and correctly segmented it. Moreover, with the visualization of the segmentation results coincident with GT in Fig. 7, it can also be inferred that the proposed method can accurately segment polyp edges. As shown in Fig. 5, we can draw similar conclusions on Kvasir-Instrument dataset. Compared with other methods, our proposed NA-SegFormer method achieved more accurate segmentation of small target.

In addition, Fig. 6 shows the segmentation results in the KvasirCapsule-SEG dataset, where the polyp tissue is very larger and shows little difference from the healthy surrounding tissue, posing a considerable challenge to precise segmentation methods. As shown in Fig. 6, Some methods, such as the UNet, Unet++, ResUnet++, SETR, SwinUNet method fails to segment the darker polyps effectively. Although ResUNet and MissFormer can segment some of the darker polyps, it still has a significant gap compared to the true labels. In contrast, the TransUnet and the proposed method get better segmentation results. If TransUnet does not use the pre-trained model, TransUnet$^{\#}$ in Fig. 6, the segmentation effect decreases significantly. To sum up, the proposed method achieved considerable performance gains in these challenging samples compared to the baseline and other remarkable models, which indicates that our method could handle complex exogenous data effectively.

To further study the computational complexity and the potential of real-time segmentation of the compared method, we used Frame-per-second (FPS) as the evaluation indicator for the computational consumption; the results are reported in Table 5. To provide a more intuitive and comprehensive comparison, we also presented a schematic diagram to show the trade-off between inference speed and accuracy for all compared methods in Fig. 8.

From the Table 5, it can be seen that our method had a moderate and satisfactory trade-off between inference speed and accuracy. More specifically, our method obtained an FPS of 125.01, ranking third among the 9 comparison methods, but the accuracy was the highest. In terms of computational cost and accuracy, our method has excellent performance in polyp segmentation, which also makes it applicable to real-time endoscopic intestinal polyp segmentation.

## Ablation study

To evaluate the impact of each module on improvement in the segmentation accuracy of colonoscopic polyps, we conducted four ablation experiments on our proposed NASegFormer network. The experiments evaluated the contribution of each module at both quantitative and qualitative levels. The training, testing, and hyperparameter settings were consistent with those described in Section "Experimental Setup".

Four models with different configurations were selected for the comparison: (1) the L1 model, which was the baseline (the basic NASegFormer network); (2) the L2 model, which added the Unified Focal Loss (UFL) on top of L1; (3) the L3 model, which added the Neighborhood Attention (NA) on top of L1; (4) the L4 model,

| Model | UNet | UNet++ | ResUNet | ResUNet++ | SETR | SwinUnet | TransUnet | MissFormer | Ours |
|---|---|---|---|---|---|---|---|---|---|
| FPS | 161.73 | 139.27 | 124.91 | 92.52 | 189.32 | 78.23 | 51.73 | 47.03 | 125.01 |

**Table 5**. Computational complexity analysis of the compared methods.

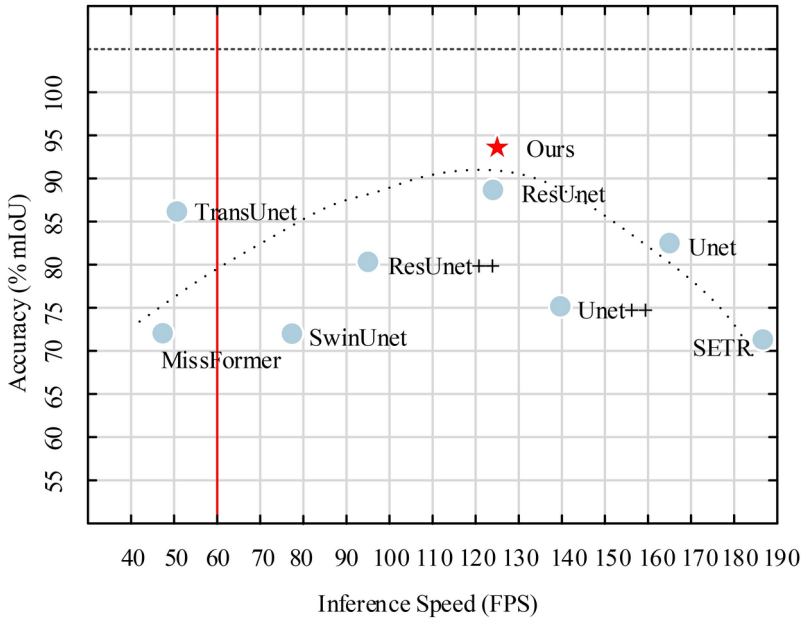


**Fig. 8**. The trade-off between inference speed and accuracy for compared models on the Kvasir-SEG test set. Red represents our model, and green represents other models.

| Models | UFL | NA | Precision | Recall | Dice | mIoU | Accuracy |
|---|---|---|---|---|---|---|---|
| L1 (baseline) | | | 83.71 | 82.12 | 82.91 | 82.45 | 94.84 |
| L2 | √ | | 73.88 | **95.12** | 83.17 | 83.08 | 95.53 |
| L3 | | √ | 83.76 | 94.13 | 88.65 | 87.97 | 96.79 |
| L4 (ours) | √ | √ | **96.14** | 92.53 | **94.30** | **93.59** | **98.26** |

**Table 6**. Impacts comparison of each module (**Bolded** parts represent better performance).

which used the NASegFormer network framework, UFL, and NA, i.e., the model proposed in this article. Table 6 presents the experimental results on Kvasir-SEG dataset from the ablation study, including five metrics.

From Table 6, the following conclusions could be drawn: First, the UFL improved in four indicators, i.e., Recall, Dice, mIoU, and Accuracy, compared to baseline, particularly a 13% improvement in Recall. This implies that UFL overcomes the category gradient bias of background class and has a positive effect on segmentation. Second, the NA mechanism contributed the most to our proposed approach, which increased the five indicators Precision, Recall, Dice, mIoU, and Accuracy by 0.05%, 12.01%, 5.74%, 5.52% and 1.95%, respectively. In particular, NA significantly improved Dice and mIoU. This is due to the fact that NA can effectively mine the semantic information of adjacent regions in our network framework and effectively improve the segmentation edge accuracy. Finally, it should be emphasized that the proposed method, that is, the L4 model, took full advantage of the each module and achieved a comprehensive improvement. Especially for Precision, Dice and mIoU compared with the baseline and the methods embedded with only one module, they all achieved significant improvement. This implies that the NA and UFL can form a good complement and play an important role in the network. Together, they can effectively capture discriminant features of polyp images, thereby improving segmentation accuracy.

To provide an interpretative analysis of our model, we generated a series of visualization results with heat maps and fine grained segmentation area presentation. Some representative examples of our model and the baseline are presented in Fig. 9. It can be seen from the experimental results that both NA and UFL have improved the baseline. More specifically, As shown in line 6 of Fig. 9, L3 can effectively improve FP (i.e., Blue region: non-polyp pixels incorrectly labeled as polyp) against the Baseline, which indicates that NA module paid more attention to adjacent features and thus reduced the misdiagnosis rate of suspected polyp tissue. Compared with Baseline, UFL can further explore the feature of rare categories of polyp medical images. A representative example was shown in row 4 and column 8 of Fig. 9; even if the proportion of polyp tissue in the whole medical image was low, UFL can still provide effective segmentation gradient guidance. This implies that UFL can effectively optimize gradients when dealing with the issue of category imbalance in polyp datasets. By taking full advantage of both NA and UFL, our proposed approach, i.e., L4 model, achieves maximum improvement. As shown in lines 5 and 6 in Fig. 9, our approach can effectively capture global and local features of polyp issue, thereby improving segmentation accuracy.

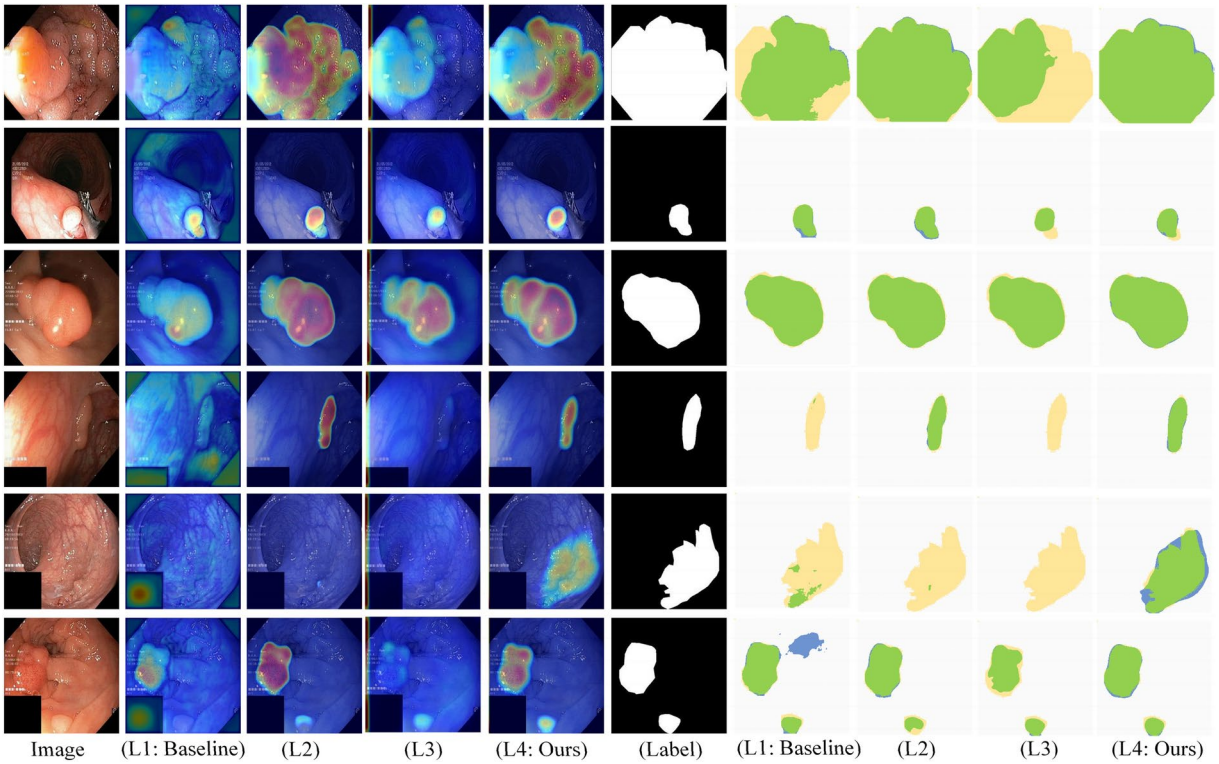


**Fig. 9**. Visualization results with heat maps and fine grained segmentation area presentation of ablation experiments in the Kvasir-SEG dataset. (In the last four columns, Green denotes TP; Blue denotes FP; Yellow denotes FN; Gray denotes TN).

## Conclusions

This paper presents a novel Transformer-based multi-level encoder-decoder architecture with a neighborhood attention mechanism and unified focal loss for colonoscopic polyp image segmentation. Specifically, a Transformer backbone network embed with multiple optimization strategy was designed to mine and exploit both local and global features that improve the performance of the entire learning system. Two key modules should be noted in our network model. One was the neighbor attention module, with that, spatial adjacent features and context information could be captured. The other was unified focal loss, which was outstanding in mining the characteristics of rare categories of polyp medical images, and can effectively guide gradient optimization and overcome the influence of data imbalances. To comprehensively evaluate the proposed method, we organized rigorous experiments on three public benchmark dataset for colonoscopic polyp segmentation. Considering the segmentation precision and time consumption comprehensively, the experimental results indicated that the proposed model achieved the best performance among the eight compared methods, in terms of the objective evaluation indicators. The ablation experiments demonstrated that the modules in our model can effectively exploit their respective advantages and support each other to further improve the segmentation performance. Based on the visual analysis of a large number of segmented samples, our method can accurately segment the edges of polyps. Even in the face of an untrained external dataset, it can also show satisfactory segmentation effect. In particular, our method achieved a distinguished trade-off between inference speed and accuracy, which are expected to have important reference values for real-time colonoscopic polyp segmentation.

Further improvements can be achieved by incorporating global context to enhance feature extraction, leading to a more accurate and efficient analysis of colorectal cancer in clinical applications.

## Data Availability

The datasets used in this paper are all publicly available. The Kvasir- SEG dataset is 544 a publicly available dataset that can be found at https://datasets.simula.no/kvasir-seg/. The KvasirCapsule-SEG da-545 taset is a publicly available dataset that can be found at Simula Datasets - KvasirCapsule SEG. The Kvasir-Instru-546 ment dataset is a publicly available dataset that can be found at https://datasets.simula.no/kvasir-instrument/.

Received: 29 April 2024; Accepted: 24 September 2024
Published online: 28 September 2024

## References

1. Siegel, R. L. *et al.* Cancer statistics, 2023 [J]. *CA Cancer J Clin* **73**(1), 17–48 (2023).

2. Krenzer, A. *et al.* Fast machine learning annotation in the medical domain: a semi-automated video annotation tool for gastroenterologists[J]. *BioMedical Engineering OnLine* **21**(1), 33 (2022).
3. Brand, M. *et al.* Development and evaluation of a deep learning model to improve the usability of polyp detection systems during interventions[J]. *United European Gastroenterology Journal* **10**(5), 477–484 (2022).
4. Wan, J., Chen, B. & Yu, Y. Polyp detection from colorectum images by using attentive YOLOv5[J]. *Diagnostics* **11**(12), 2264 (2021).
5. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings,* part III 18 (pp. 234–241). Springer International Publishing. (2015)
6. Zhang, R., Li, G., Li, Z., Cui, S., Qian, D., & Yu, Y. Adaptive context selection for polyp segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings,* Part VI 23 (pp. 253–262). Springer International Publishing. (2020).
7. Strudel, R., Garcia, R., Laptev, I., & Schmid, C. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 7262–7272). (2021).
8. Cao H, Wang Y, Chen J, et al. Swin-unet: Unet-like pure transformer for medical image segmentation[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 205–218.
9. Chen, J. *et al.* TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers[J]. *Medical Image Analysis* **97**, 103280 (2024).
10. Zheng, S. *et al.* Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers[C]//Computer Vision and Pattern Recognition. *IEEE* https://doi.org/10.1109/CVPR46437.2021.00681 (2021).
11. Huang, X. *et al.* Missformer: An effective transformer for 2d medical image segmentation[J]. *IEEE Transactions on Medical Imaging* **42**(5), 1484–1494 (2022).
12. Brand, M. *et al.* Frame-by-frame analysis of a commercially available artificial intelligence polyp detection system in full-length colonoscopies[J]. *Digestion* **103**(5), 378–385 (2022).
13. Long, J., Shelhamer, E., & Darrell, T. Fully convolutional networks for semantic segmentation. In *IEEE conference on computer vision and pattern recognition*. pp. 3431–3440 (2015).
14. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence* **39**, 640–651. https://doi.org/10.1109/TPAMI.2016.2572683 (2017).
15. Liu, W., Rabinovich, A., & Berg, A. C. Parsenet: Looking wider to see better. *arXiv preprint* arXiv:1506.04579. (2015).
16. Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint* arXiv:1412.7062. (2014).
17. Badrinarayanan, V., Kendall, A. & Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence* **39**(12), 2481–2495 (2017).
18. Yuan, Y., Chen, X., & Wang, J. Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings,* Part VI 16 (pp. 173–190). Springer International Publishing. (2020).
19. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., & Lu, H. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3146–3154). (2019).
20. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* arXiv:1810.04805. (2018).
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, **30** (2017).
22. Xie, E. et al. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems* **34**, 12077–12090 (2021).
23. Zheng, S, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6881–6890). (2021).
24. Wang, J., Wei, L., Wang, L., Zhou, Q., Zhu, L., & Qin, J. Boundary-aware transformers for skin lesion segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings,* Part I 24 (pp. 206–216). Springer International Publishing. (2021).
25. Lee, J. H., Kim, D. H., Jeong, S. N. & Choi, S. H. Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. *Journal of dentistry.* **77**, 106–111 (2018).
26. Lee, J. H., Han, S. S., Kim, Y. H., Lee, C. & Kim, I. Application of a fully deep convolutional neural network to the automation of tooth segmentation on panoramic radiographs. *Oral surgery, oral medicine, oral pathology and oral radiology.* **129**(6), 635–642 (2020).
27. Shen, Z., Yang, H., Zhang, Z., & Zheng, S. Automated kidney tumor segmentation with convolution and transformer network. In *International Challenge on Kidney and Kidney Tumor Segmentation* (pp. 1–12). Cham: Springer International Publishing. (2021).
28. Zhang, X. *et al.* Real-time gastric polyp detection using convolutional neural networks. *PloS one* **14**(3), e0214133 (2019).
29. Misawa, M. *et al.* Artificial intelligence-assisted polyp detection for colonoscopy: initial experience. *Gastroenterology* **154**(8), 2027–2029 (2018).
30. Qadir H A, Shin Y, Bergsland J, et al. Accurate Real-time Polyp Detection in Videos from Concatenation of Latent Features Extracted from Consecutive Frames[C]//2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2022: 2461–2466.
31. Krenzer A, Hekalo A, Puppe F. Endoscopic Detection And Segmentation Of Gastroenterological Diseases With Deep Convolutional Neural Networks[C]//EndoCV@ ISBI. 2020: 58–63.
32. Guo, X. *et al.* Non-equivalent images and pixels: Confidence-aware resampling with meta-learning mixup for polyp segmentation[J]. *Medical image analysis* **78**, 102394 (2022).
33. Akbari, M., et al. Polyp segmentation in colonoscopy images using fully convolutional network. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 69–72). IEEE. (2018).
34. Qadir H A, Shin Y, Solhusvik J, et al. Polyp detection and segmentation using mask R-CNN: Does a deeper feature extractor CNN always perform better?[C]//2019 13th International Symposium on Medical Information and Communication Technology (ISMICT). IEEE, 2019: 1–6.
35. Jha, D., et al. Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE international symposium on multimedia (ISM)* (pp. 225–2255). IEEE. (2019).
36. Fan, D. P., et al. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 263–273). Cham: Springer International Publishing. (2020).
37. Galdran, A., Carneiro, G., & Ballester, M. A. G. Double encoder-decoder networks for gastrointestinal polyp segmentation. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings,* Part I (pp. 293–307). Springer International Publishing. (2021).
38. Mahmud, T., Paul, B. & Fattah, S. A. PolypSegNet: A modified encoder-decoder architecture for automated polyp segmentation from colonoscopy images. *Computers in biology and medicine* **128**, 104119 (2021).
39. Chen, J., et al. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint* arXiv:2102.04306. (2021).

40. Xiao B, Hu J, Li W, et al. CTNet: Contrastive Transformer Network for Polyp Segmentation[J]. IEEE Transactions on Cybernetics, 2024.
41. Krenzer, A. *et al.* Automated classification of polyps using deep learning architectures and few-shot learning[J]. *BMC Medical Imaging* **23**(1), 59 (2023).
42. Liu, X. & Song, S. Attention combined pyramid vision transformer for polyp segmentation. *Biomedical Signal Processing and Control* **89**, 105792 (2024).
43. Jha, D., N. K, Ali, S., et al. Nanonet: Real-time polyp segmentation in video capsule endoscopy and colonoscopy. In *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)* (pp. 37–43). IEEE. (2021).
44. Lin, L. *et al.* Polyp-LVT: Polyp segmentation with lightweight vision transformers. *Knowledge-Based Systems* **300**, 112181 (2024).
45. Krenzer, A., Banck, M., Makowski, et al. A Real-Time Polyp-Detection System with Clinical Application in Colonoscopy Using Deep Convolutional Neural Networks. J. Imaging 2023, 9, 26. https://doi.org/10.3390/jimaging9020026
46. Hassani, A., Walton, S., Li, J., Li, S., & Shi, H. Neighborhood attention transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6185–6194). (2023).
47. Yeung, M., Sala, E., Schönlieb, C. B. & Rundo, L. Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Computerized Medical Imaging and Graphics* **95**, 102026 (2022).
48. Loshchilov, I., & Hutter, F. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101. (2017).
49. Vakili, M., Ghamsari, M., & Rezaei, M. Performance analysis and comparison of machine and deep learning algorithms for IoT data classification. *arXiv preprint* arXiv:2001.09636. (2020).
50. Jha, D., et al. Kvasir-seg: A segmented polyp dataset. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings,* Part II 26 (pp. 451–462). Springer International Publishing. (2020).
51. P. H. Smedsrud et al., Kvasir-capsule, a video capsule endoscopy dataset, *Springer Nature Scientific Data*, 2021.
52. Jha, D., et al. Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy. In *MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings,* Part II 27 (pp. 218–229). Springer International Publishing. (2021).
53. Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N. & Liang, J. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging* **39**(6), 1856–1867 (2019).



## Acknowledgements
The author thanks the reviewers and other scholars for providing precious suggestion on this manuscript. And we also thank KetengEdit (www.ketengedit.com) for its linguistic assistance during the preparation of this manuscript.


## Author contributions
Conceptualization, D.L., C.L., H.S and S.G.; methodology, D.L. and C.L.; software, C.L., H.S. and D.L.; validation, C.L., D.L. and S.G.; formal analysis, D.L., C.L., H.S. and S.G.; investigation, D.L., C.L., H.S. and S.G.; resources, S.G. and D.L.; data curation, H.S. and C.L.; writing—original draft preparation, D.L. and C.L.; writing—review and editing, D.L., C.L. and S.G.; visualization, C.L.; supervision, D.L. and S.G.; project administration, D.L. and S.G.; funding acquisition, S.G. and D.L. All authors have read and agreed to the published version of the manuscript.


## Funding
This research was funded by Scientific Research Fund of Hunan Provincial Education Department (No. 23A0588), Natural Science Foundation of Hunan Province (No. 2023JJ50392) and Aid Program for Science and Technology Innovative Research Team in Higher Educational Institutions of Hunan Province.


## Declarations

## Competing interests
The authors declare no competing interests.

## Additional information
**Correspondence** and requests for materials should be addressed to S.G.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.