BMC Medical Imaging

RESEARCH

Open Access

DCATNet: polyp segmentation with deformable convolution and contextual-aware attention network

Zenan Wang[1], Tianshu Li[1], Ming Liu[2], Jue Jiang[3] and Xinjuan Liu[1*]

Abstract

Polyp segmentation is crucial in computer-aided diagnosis but remains challenging due to the complexity of medical images and anatomical variations. Current state-of-the-art methods struggle with accurate polyp segmentation due to the variability in size, shape, and texture. These factors make boundary detection challenging, often resulting in incomplete or inaccurate segmentation. To address these challenges, we propose DCATNet, a novel deep learning architecture specifically designed for polyp segmentation. DCATNet is a U-shaped network that combines ResNetV2-50 as an encoder for capturing local features and a Transformer for modeling long-range dependencies. It integrates three key components: the Geometry Attention Module (GAM), the Contextual Attention Gate (CAG), and the Multi-scale Feature Extraction (MSFE) block. We evaluated DCATNet on five public datasets. On Kvasir-SEG and CVC-ClinicDB, the model achieved mean dice scores of 0.9351 and 0.9444, respectively, outperforming previous state-of-the-art (SOTA) methods. Cross-validation further demonstrated its superior generalization capability. Ablation studies confirmed the effectiveness of each component in DCATNet. Integrating GAM, CAG, and MSFE effectively improves feature representation and fusion, leading to precise and reliable segmentation results. These findings underscore DCATNet's potential for clinical application and can be used for a wide range of medical image segmentation tasks.

Keywords Transformer, Deformable attention, Colorectal Polyp, polyp segmentation, Deep learning

Introduction

Colorectal cancer (CRC) is one of the most common cancers worldwide and has the second-highest mortality rate [1]. Early detection and removal of polyps are critical to preventing the progression to cancer. However, the effectiveness of colonoscopy is heavily dependent on the experience of the endoscopist. Missed polyp detection remains a major clinical concern. Conventional colonoscopy still has a relatively high rate of missed detection for colorectal polyps and adenomas, thereby posing a risk of interval cancer (colorectal cancer that occurs between a colonoscopy with normal results or after all polyps have been removed and the next colonoscopy). Studies have shown that up to 20–30% of polyps can be

*Correspondence:
Xinjuan Liu
liuxinjuan@mail.ccmu.edu.cn
[1]Department of Gastroenterology, Beijing Chaoyang Hospital, The Third Clinical Medical College of Capital Medical University, Beijing, China
[2]Hunan Key Laboratory of Nonferrous Resources and Geological Hazard Exploration, Changsha, China
[3]Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York City, NY, USA

missed, particularly smaller or flat polyps [2, 3]. Previous studies have shown that 58% of interval cancers occur due to inadequate initial colonoscopy leading to missed lesions, and 89% of interval cancers can be prevented [4]. The adenoma detection rate is significantly negatively correlated with the risk of interval cancer and its associated mortality [5]. Additionally, missed polyps can delay the diagnosis of colorectal cancer, leading to a significantly lower survival rate. Research indicates that for every 1% increase in polyp detection, the incidence of bowel cancer decreases by 3% [6]. Computer-aided diagnostic (CAD) systems for automatic polyp segmentation can accurately locate and segment polyps. These systems improve the efficiency and accuracy of detection while reducing errors caused by manual intervention [7]. However, the diverse shapes and sizes of polyps, along with their blurred edges, and polyps can be confused with folds, making them difficult to distinguish from normal tissue. Addressing these challenges remains difficult for existing methods.

In recent years, Convolutional Neural Networks (CNNs) have shown strong representation capabilities and achieved great success in medical image segmentation. Among these, the classic U-Net [8] with its symmetric encoder-decoder structure has performed remarkably well in medical image semantic segmentation. The encoder extracts multi-level features, while skip connections transfer complex and rich features from the encoder to the decoder. The decoder then generates the final segmentation predictions. Building on this success, many new models have been developed for medical image and polyp segmentation, such as ResUNet++ [9], AttUNet [10], U^2-Net [11], M^2SNet [12], and U-Net++ [13]. These models improve several aspects of U-Net. However, the structural limitations of the convolutional operator make it hard for pure CNN-based models to capture long-range dependencies. These dependencies are crucial for accurately locating lesion areas and boundaries.

With great success in Natural Language Processing (NLP), Vision Transformer (ViT) [14] has achieved state-of-the-art performance in computer vision tasks, such as image classification, detection, and segmentation. In contrast to CNNs, Transformer models global relationships between pixels, and can effectively capture long-range dependency. However, they lack spatial sensing bias, which limits their ability to extract local features, which is crucial for analyzing complex medical images. CNNs are good at capturing local details, while Transformers are effective at extracting global information. To leverage both advantages, many researchers have combined Transformers with CNNs. TransUNet [15] integrates Transformer blocks into the U-Net structure to model long-range dependencies. However, TransUNet is not proficient in capturing multi-scale features and modeling geometric features, which are very important for polyp segmentation tasks. Additionally, TransUNet ignores the feature semantic differences between various feature extraction mechanisms. Unlike TransUNet, the Polyp-Pvt [16] uses the pyramid Transformers as the encoder, it integrates three submodules to collect the semantic information from high-level features, capture contextual information from low-level features, and fuse the high-level and low-level features with non-local operations. However, it incorporates channel and spatial attention which are not good at capturing the geometric features of the polyp boundaries. Furthermore, the design of Polyp-Pvt is not well-suited for multi-scale feature extraction and fusion. TransFuse [17] uses a dual-branch architecture with CNN and Transformer in parallel, fusing local and global features at the same stage with a BiFusion module. DSTransUNet [18] employs a dual-branch Swin Transformer encoder to capture multi-scale features and model global dependencies. ColonFormer [19] combines a lightweight Transformer encoder with a CNN decoder to capture global semantic information at multiple scales, improving segmentation accuracy. Segformer [20] processes features at different scales and depths separately, then combines them using a parallel multi-stage feature aggregation algorithm. SSFormer [21] uses a Transformer as an encoder and CNN as a decoder, enabling progressive prediction. Additionally, several hybrid CNN-Transformer models have demonstrated high performance [22–24]. However, existing hybrid models often ignore the semantic differences in features and polyp geometric information, which may lead to inaccurate segmentation results, still leaving much room for improvement in polyp and medical segmentation.

One of the major problems is that the semantic features extracted from the CNNs often contain noises, which prevents the model from improving segmentation performance. The attention mechanism helps neural networks focus on important parts of their inputs. This improves the model's accuracy and efficiency, making it a key component in many models. Researchers have applied attention mechanisms to enhance polyp segmentation performance. AttUNet [10] introduces attention gates to combine high-level and low-level features, focusing on target regions while ignoring irrelevant areas. Dual Attention Network (DANet) [25] uses the channel attention module and position attention module to process features parallelly with channel and spatial attention blocks. PraNet [26] uses reverse attention modules to refine key polyp regions, improving segmentation accuracy. CAFE-Net [27] proposes a cross-attention decoder module (CADM) to retain early features and recover fine details. These attention mechanisms have been proven effective in improving polyp segmentation performance.

TANet [28] proposed a triple attention module to enhance the segmentation performance, which adaptively selects the optimal scale feature from multi-scale features.

The ambiguous boundary of some polyps results in over- or under-segmentation, so the accuracy of the polyp segmentation cannot be guaranteed, and the variety of the polyps' sizes and shapes makes it more difficult to segment from the surroundings. To address these challenges, we propose DCATNet, a novel deep-learning architecture for polyp segmentation. DCATNet follows a U-shaped architecture. It uses ResNetV2-50 [29] as the encoder to capture multi-layer local information, while a Transformer [14] captures global information and models long-range dependencies. Additionally, DCATNet incorporates two attention modules, the **G**eometry **A**ttention **M**odule (GAM) and **C**ontextual **A**ttention **G**ate (CAG) module, and a **M**ulti-**S**cale **F**eature **E**xtraction and fusion model. The GAM module captures spatial and geometric information, such as shapes and boundaries. The CAG module integrates contextual and semantic information, reducing the semantic gap between the encoder and decoder. The MSFE module extracts and fuses multi-scale features to better capture polyps of varying sizes. In summary, our contributions are as follows:

1. We propose a novel deep-learning architecture DCATNet for polyp segmentation.
2. Geometry Attention Modules (GAM): Building on deformable convolution, integrating with residual connection, the GAM can be used to dynamically learn the spatial and geometric features from input. This enables the model to extract more discriminative features, which are crucial for accurate segmentation.
3. To fuse features from the encoder and decoder and reduce the semantic gap, we design a Contextual Aggregation Gate (CAG) module. It uses contextual features to guide low-level features, helping the model focus on the important areas of the polyp. This enables accurate segmentation of complex polyps.
4. We employ MSFE as the decoder block to better capture and fuse the multi-scale features.
5. Extensive experiments on five benchmark datasets demonstrate that DCATNet consistently outperforms existing state-of-the-art methods across multiple metrics. Ablation studies further validate the contribution of each component in DCATNet.

Method

Overall network architecture

In this paper, we propose a novel approach for polyp segmentation based on the U-Net framework. The model architecture includes a ResNetV2-50-based encoder, a Transformer module, and an MSFE decoder based on four stages of residual U-block (RSU) modules [11]. The model incorporates the Geometry Attention Module (GAM) and Contextual Attention Gate (CAG) module for feature extraction and fusion. The Transformer bridges the encoder and decoder, capturing long-range dependencies and contextual information. The overall structure of the model is shown in Fig. 1.

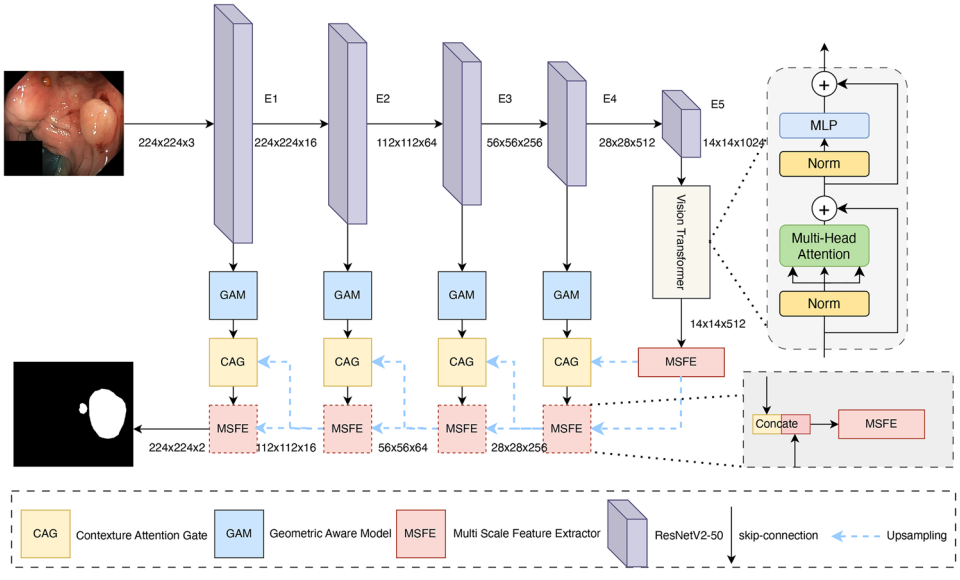

Fig. 1 Overall structure of the proposed model. GAM is employed to capture additional spatial features from the decoder, maintaining an equal number of input and output channels. CAG aggregates features from both the encoder and decoder to reduce the semantic gap. MSFE serves as the decoder for multi-scale feature extraction and fusion. 12 Transformer layers are incorporated in this model

The encoder consists of five convolutional stages progressively downsampling the input image to extract hierarchical features. The final encoder layer feeds into a 12-layer Transformer module. The Transformer's output initializes the decoder, which upsamples features to match the spatial dimensions of corresponding encoder layers. Encoder and decoder are connected by a skip-connection with a GAM module. After that, features from the output of GAM, and the upsampled decoder are fused with the CAG module, which employs contextual attention to integrate features while emphasizing relevant information and suppressing noise. The fused features and the previous decoder features are passed through the MSFE module for multi-scale feature extraction and fusion, enriching feature representation. This process repeats for each decoder layer, refining feature maps and preserving spatial information, ultimately producing a segmentation map with precise and detailed delineation of target structures. By combining ResNetV2-50 for feature extraction, Transformers for long-range dependency capture, and GAM, CAG, and MSFE modules for effective feature fusion and refinement, DCATNet achieves precise and robust segmentation of complex medical images.

Geometry attention module

The variety in geometric shapes and sizes poses challenges for polyp detection and segmentation, which prevents the model's ability to focus on key details and contextual features. This issue is compounded by the varying degrees of noise present in the multi-layer features extracted by the encoder. To address these challenges in complex environments, we introduce a GAM module to enhance feature extraction capabilities for colon polyp segmentation. The detailed structure of GAM is shown in Fig. 2. Unlike standard skip connections, which directly transfer low-level features from the encoder to the decoder without any modulation, this module based on deformable convolutions can dynamically learn spatial and geometric features. This allows the module to focus on critical boundary and shape information and is specifically designed for integration into the skip connections between the encoder and decoder stages.

The GAM consists of three distinct branches: a convolutional branch with a kernel size of 1×1, a convolutional branch with a kernel size of 3×3, and a deformable convolutional branch. The input features x are simultaneously passed through these three branches in parallel. The deformable convolutional branch is designed to capture boundary and geometric features, sigmoid activation function is applied to generate an activation map, which is then used to re-weight the outputs of the other two branches. Finally, the re-weighted features are combined using element-wise addition to produce the final output features x_{out}. This process is mathematically described by the following equations:

$$x_1 = Conv_{1\times1}(x) \tag{1}$$

$$x_2 = DCN(x) \tag{2}$$

$$x_3 = Conv_{3\times3}(x) \tag{3}$$

$$x_1' = x_1 \otimes \sigma(x_2) \tag{4}$$

$$x_3' = x_3 \otimes \sigma(x_2) \tag{5}$$

$$x_{out} = Conv_{1\times1}(x_1' \oplus x_3') \tag{6}$$

Where $Conv_{1\times1}$ stands for the convolutional operation with a kernel size of 1×1, $Conv_{3\times3}$ stands for the convolutional operation with a kernel size of 3×3, σ stands for sigmoid active function, $\otimes$ represents Hadamard product, $\oplus$ means element-wise addition, DCN stands for the deformable convolutions.

Contextual attention gate

One of the major concerns in polyp segmentation is the similarity in texture between polyps and surrounding mucosal tissue, making segmentation challenging. High-level features often contain richer semantic

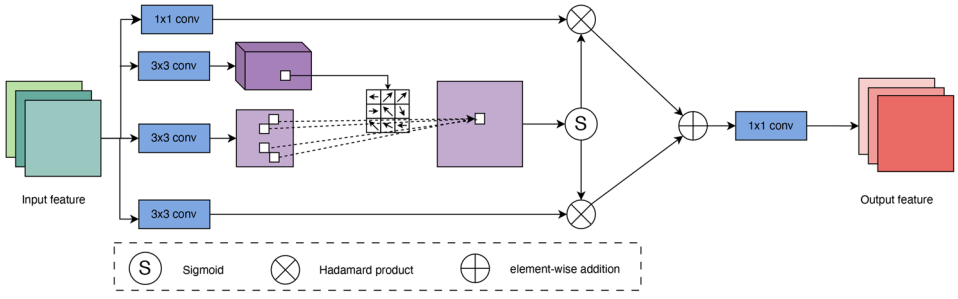

Fig. 2 Geometry Attention Module structure. The GAM module consists of three parallel branches: a 1×1 convolution, a 3×3 convolution, and a deformable convolution. The input features are passed through all branches simultaneously. The deformable convolution captures boundary features, and a sigmoid activation generates an activation map to re-weight the outputs of the other branches

information, while low-level features have higher resolution and contains more detailed information, which is especially important for segmenting small polyps and edges. However, these details often introduce noise, and directly fusing low-level with high-level features can lead to redundancy, inconsistency, and an increased semantic gap between the encoder and decoder. To overcome this issue, we introduce the CAG module to fuse the features extracted from both the encoder and decoder pathways through a context-aware attention mechanism, ensuring the model focuses on the important features of the polyp. This module dynamically modulates low-level features based on contextual information, improving feature alignment and coherence. The CAG module helps the model capture and emphasize critical anatomical details. It also suppresses irrelevant information and focuses on relevant regions, improving the model's capability to delineate complex and varied anatomical structures.

The CAG module processes high-level features x_h and low-level features x_l. It first applies a convolutional layer with a kernel size of 1×1 to reduce the number of channels. This results in the outputs x'_h and x'_l. These two features are then concatenated, and a sigmoid function is applied to generate the weight map σ. The weight map α is used to re-weight x'_h and x'_l. Finally, the re-weighted features are combined using element-wise addition to produce the final output features. The detailed structure of CAG is shown in Fig. 3 The module can be defined as follows.

$$x'_h = Conv_{1\times1}(x_h) \tag{7}$$

$$x'_l = Conv_{1\times1}(x_l) \tag{8}$$

$$x' = Conv_{1\times1}([x'_h, x'_l]) \tag{9}$$

$$\alpha = \sigma(Conv_{1\times1}(x')) \tag{10}$$

$$x_{out} = Conv_{1\times1}(Conv_{1\times1}(\alpha \otimes x'_h) \oplus Conv_{1\times1}(\alpha \otimes x'_l)) \tag{11}$$

where, $Conv_{1\times1}$ means convolution operation with kernel size of 1×1, $[\cdot]$ is the concatenation operation. $0 < \alpha < 1$ denotes the attention score, σ is the sigmoid activation function, $\otimes$ represents Hadamard product, $\oplus$ means element-wise addition.

Multi-scale features extraction and fusion

Polyps vary in shape and size and often have a texture similar to their surroundings, making detection and segmentation challenging. Therefore, it is essential for the model to generalize across these variations. To address scale variation, capturing multi-scale contextual features is crucial. This improves the model's robustness and enhances its contextual understanding. Inspired by the U^2-Net [11] architecture, we use a modified four-stage U-Net block as the decoder to capture multi-scale features effectively. Each stage of the U-Block consists of standard convolutional layers, batch normalization, and ReLU activation functions. Unlike the original U^2-Net, we omit dilated convolutions, keeping them only in the last encoder stage to simplify the structure and focus on key feature extraction. The U-Block captures features from input maps with varying spatial resolutions. These features are obtained from progressively downsampled maps, allowing the network to aggregate more abstract information. The features are then upsampled, concatenated, and convolved to reconstruct high-resolution maps, which helps preserve fine details that may be lost with large-scale upsampling. Finally, a residual connection combines local and multi-scale features to enhance the final output. This module's design ensures that the network can effectively handle varying feature scales and spatial hierarchies, enhancing the decoder's ability to capture multi-scale features. The detailed structure of MSFE is shown in Fig. 4.

Loss function

The hybrid loss of Binary cross-entropy $\mathcal{L}_{bce}$ and dice loss $\mathcal{L}_{dice}$ is used in the proposed method. As defined in

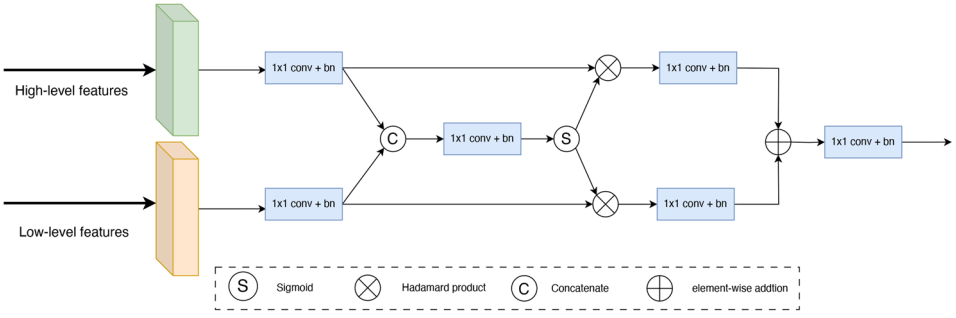

Fig. 3 Contextual Attention Gate Module structure. The CAG module fuses high-level and low-level features through a context-aware attention mechanism. High-level and low-level features are first processed with 1×1 convolutions, then concatenated. The sigmoid function generates a weight map to re-weight the features. The final output is generated by element-wise addition

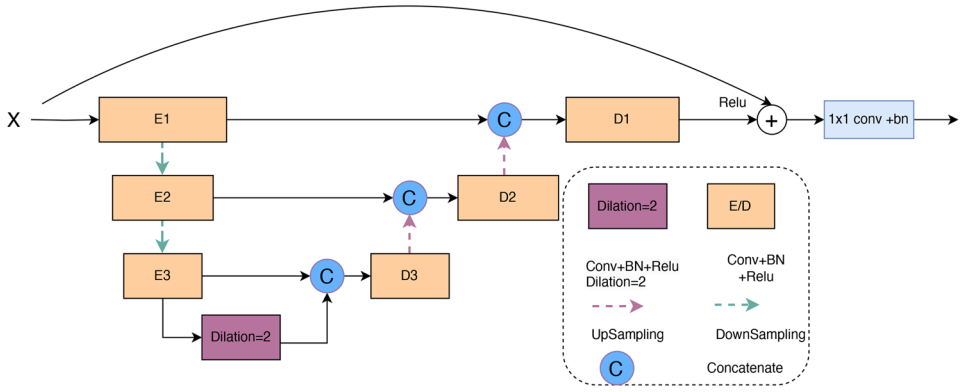

Fig. 4 The multi-scale feature extraction module consists of an encoder and decoder, each built with standard convolutional blocks, batch normalization, and ReLU activation. The final stage of the encoder uses dilation = 2 to capture global information

Eq. 14, where y denotes the ground truth and $\hat{y}$ denotes the prediction.

$$\mathcal{L}_{bce} = (y-1)log(1-\hat{y}) - ylog\hat{y} \tag{12}$$

$$\mathcal{L}_{dice} = 1 - \frac{2y\hat{y}+1}{y+\hat{y}+1} \tag{13}$$

$$\mathcal{L}_{decoder} = \mathcal{L}_{bce} + \mathcal{L}_{dice} \tag{14}$$

Experiment

Datasets

We employ five publicly available datasets, commonly utilized in biomedical image segmentation, to evaluate our method. The primary reason for selecting these diverse imaging modality datasets is to assess the performance and robustness of the proposed method comprehensively.

Kvasir-SEG [30] comprising 1,000 images and their corresponding ground truth masks. The dataset features a wide range of resolutions, from 332×487 to 1920×1072 pixels.

CVC-ColonDB [31] includes 380 images, each accompanied by polyp masks derived from 13 polyp video sequences from 13 patients. The images are uniformly sized at a resolution of 574×500 pixels. All the data in this dataset are used for cross-validation.

CVC-ClinicDB [32] contains 612 frames selected from 29 various colonoscopy videos. The images have a resolution of 384×288 pixels.

ETIS [33] This dataset contains 192 polyp images, along with their annotation. Each image in this dataset maintains a consistent resolution of 1225×966 pixels.

CVC-300 [34] This dataset is the test set from EndoScene [34] and contains 60 polyp images, each with a size of 500×574.

Evaluation metrics

In this study, we use four standard evaluation metrics to better evaluate the segmentation performance of our model. They are commonly used in the medical image segmentation field: mean Dice Coefficient (mDice), mean Intersection over Union (mIoU), Recall, and Precision. Dice quantifies the similarity between the ground truth and the prediction, and IoU calculates the overlap between the ground truth and the prediction image.

$$\begin{array}{llll} \text{Dice} & = \frac{2\times TP}{2\times TP+FP+FN} & \text{IoU} & = \frac{TP}{TP+FP+FN} \\ \text{Recall} & = \frac{TP}{TP+FN} & \text{Precision} & = \frac{TP}{TP+FP} \end{array} \tag{15}$$

Where TP, TN, FN, and FP represent truth positive, truth negative, false negative, and false positive respectively.

Implementation details

Pytorch framework [35] is used to implement the model. The model is trained with an SGD optimizer with momentum of 0.9 and weight decay of base learning rate of $1e^{-4}$. Learning rate schedule is defined as the formula $lr = lr_0(1 - \frac{iter}{max\_iter})^{0.9}$, where $iter$ and max_iter denote current iteration step and total iteration steps respectively. The experiments are conducted on an NVIDIA A100 Tensor Core GPU with 32GB memory. We resized all the images to 224×224 to reduce computational complexity and improve training efficiency. Furthermore, to prevent overfitting and mitigate the potential impact of biases in these data, we utilize various data augmentation strategies such as random horizontal and vertical flipping and random rotation.

Results

In this section, we present the results of the proposed method and compare them with other methods, including U-Net [8], U-Net++ [13], ResUNet++ [9], HarDNet-MSEG [36], TransUNet [15], UTNet [37], U^2-Net [11], M^2SNet [12], PGCF [7], and CoinNet [38]. Among

datasets, Kvasir-SEG and CVC-ClinicDB are used to train and test the learning ability, while CVC-ColonDB, ETIS, and CVC-300 are used to test the generalization ability. For a fair comparison, we use the same training-testing dataset and reproduce the experimental results with a unified training method. The best results in all the tables are highlighted in **bold**.

Learning ability

To evaluate the model's learning ability, we used 900 images from Kvasir-SEG and 550 images from CVC-ClinicDB for separate training. The remaining images from these two datasets were reserved for testing. The performance of different segmentation methods on Kvasir-SEG and CVC-ClinicDB is summarized in Table 1. On the Kvasir-SEG dataset, DCATNet performed better than other methods in terms of mDice, mIoU, and Precision. Compared to traditional U-Net and its variants, DCATNet showed significant improvement. Specifically, compared with TransUNet [15], DCATNet increased mDice, mIoU, and Precision by 1.34%, 2.51%, and 3.73%, respectively, though it had slightly lower Recall. The model's effectiveness was also tested on the CVC-ClinicDB dataset. DCATNet achieved excellent results across mDice and mIoU. Where, its Precision was slightly lower than M^2SNet [12], and its Recall was lower than PGCF [7]. Compared with TransUNet [15], DCATNet improved all four metrics. Specifically, mDice, mIoU, Recall, and Precision increased by 2.02%, 3.84%, 3.65%, and 0.28%, respectively. This analysis shows that DCATNet significantly improves segmentation performance. Additionally, Figs. 7 and 6 present the prediction distributions on the test datasets, further demonstrating the stability of our method. Qualitative results are shown in Fig. 5. From these, we can see that for the first three flat, small polyps, our method outperforms the others. Specifically, in the first case, only our method produces the correct results. These results indicate that other methods lose details and sometimes produce incorrect outputs. In contrast, DCATNet predictions are closer to the ground truth, reduce false segmentations, and have clear boundaries.

Generalization ability

To study the model's effectiveness further, we performed a cross-validation study using four public datasets. The models were trained on Kvasir-SEG and CVC-ClinicDB with 1450 images and tested on all five datasets. The results are shown in Table 2. Our proposed method outperformed the other methods and achieved the highest scores across all datasets. On the Kvasir-SEG dataset, DCATNet reached a mDice score of 0.9266, which was higher than TransUNet's 0.9187. Compared to TransUNet [15], our method improved the mDice score by 0.86%, 2.22%, 6.64%, 3.21%, and 4.2% on the five datasets. DCATNet also showed strong generalization on unseen datasets like ColonDB, ETIS, and CVC-300 datasets. It achieved a mDice score of 0.7872 on ColonDB, 0.8511 on ETIS, and 0.9064 on CVC-300, these results are significantly better than TransUNet [15]. While HarDNet-MSEG [36], M^2SNet [12] and PGCF [7] performed well on known datasets (Kvasir and ClinicDB), their performance on unseen datasets (ColonDB, ETIS and CVC-300) was relatively weak. These results demonstrate the robustness and generalizability of DCATNet across different datasets.

Ablation study

To evaluate the effectiveness of each component in our proposed DCATNet, we conducted an ablation study on two public datasets, Kvasir-SEG and CVC-ClinicDB. The results are shown in Table 3. We used TransUNet [15] as the baseline model and added each component progressively to assess their contributions. Compared to the baseline, our method improved the mDice and mIoU

Table 1 Quantitative evaluation of segmentation performance was conducted on the Kvasir-SEG and CVC-ClinicDB datasets. The model was trained and tested independently on each dataset. Performance was assessed using four metrics: mDice, mIoU, Recall, and Precision. Higher values, indicated in bold, represent the best results

Methods	Kvasir-SEG				CVC-ClinicDB			
	mDice	mIoU	Recall	Precision	mDice	mIoU	Recall	Precision
U-Net [8]	0.7696	0.6254	0.7390	0.8027	0.8221	0.6980	0.7729	0.8779
U-Net++ [13]	0.7806	0.6401	0.7759	0.7853	0.8520	0.7421	0.7964	0.9159
ResUNet++ [9]	0.7575	0.6097	0.7264	0.7914	0.8414	0.7263	0.7833	0.9088
HarDNet [36]	0.8860	0.7954	0.8652	0.9078	0.9081	0.8318	0.8536	0.9701
U^2-Net [11]	0.8778	0.7823	0.8612	0.8951	0.9111	0.8368	0.8696	0.9568
M^2SNet [12]	0.8872	0.7973	0.8276	0.9561	0.7862	0.6478	0.6514	**0.9914**
TransUNet [15]	0.9227	0.8566	**0.9189**	0.9266	0.9257	0.8617	0.8813	0.9748
UTNet [37]	0.8278	0.7063	0.7781	0.8844	0.8466	0.7340	0.7792	0.9266
PGCF [7]	0.9261	0.8623	0.9157	0.9366	0.9378	0.8830	**0.9436**	0.9321
CoinNet [38]	0.8696	0.8004	0.8751	0.9010	0.8739	0.8022	0.8857	0.8712
DCATNet	**0.9351**	**0.8781**	0.9103	**0.9612**	**0.9444**	**0.8948**	0.9135	0.9775

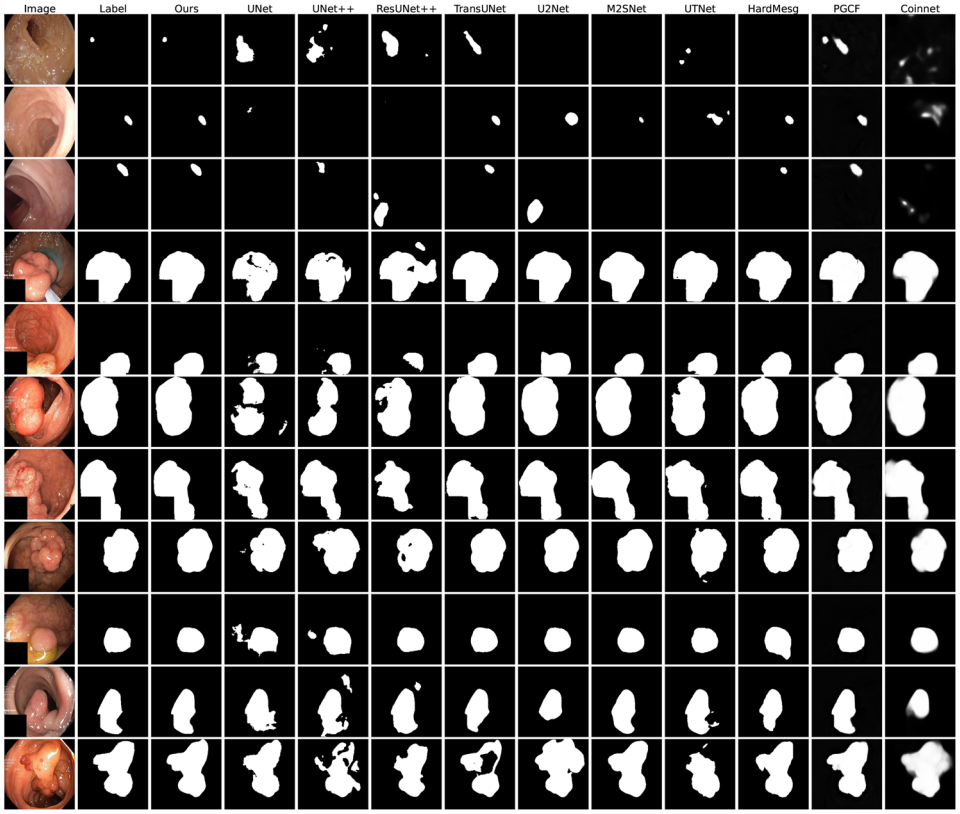

Fig. 5 Visualization of segmentation results comparing our model with ten state-of-the-art methods, including U-Net [8], U-Net++ [13], ResUNet++ [9], HarDNet-MSEG [36], TransUNet [15], UTNet [37], U^2-Net [11], M^2SNet [12], PGCF [7], and CoinNet [38]. The first three cases are from the ETIS dataset, showcasing our model's ability to accurately segment flat and small polyps. Notably, in the first case, most models fail or generate incorrect segmentations, while our method successfully predicts the correct result. The remaining cases are from the Kvasir-SEG dataset, demonstrating our model's effectiveness in segmenting large polyps and those with irregular boundaries

scores by 1.34%, 2.51%, 2.02%, and 3.84% on the two datasets. Adding the GAM module significantly improved performance, increasing the scores by 0.99%, 1.84%, 1.31%, and 2.46%. This shows that the GAM module helps the network learn and combine features better at different stages. Furthermore, we conducted an additional ablation study by replacing deformable convolutions with standard convolutional operations. The results are presented in the second-to-last row of Table 3. As shown, the performance of standard convolutions decreased by 0.89% in mDice and 1.6% in mIoU for CVC-ClinicDB, and by 1.1% in mDice and 2% in mIoU for Kvasir-SEG, indicating a clear performance drop. Next, we added the CAG module to improve feature fusion between the encoder and decoder. This further increased the scores by 0.63%, 1.17%, 1.21%, and 2.27%, proving its effectiveness in refining feature integration. Finally, replacing the standard convolutional block in the decoder with the MSFE structure improved the scores by 0.42%, 0.78%, 1.47%, and 2.76%. The MSFE module captures multi-scale features and improves spatial dependencies, which greatly enhances the overall performance. This study shows that every component plays an important role in improving DCATNet's performance.

Model complexity and inference time comparison

Table 4 compares model complexity and efficiency across different methods, with parameters (M), floating point operations (FLOPs(G)), and the inference time is measured by frames per second (FPS). Our model has the highest parameter count, primarily due to the 12-layer Transformer and deformable convolutions in the GAM module. In future work, we will focus on optimizing the network architecture by reducing computational complexity and improving segmentation performance.

Discussion

Automatic polyp segmentation is critical in the clinical diagnosis. Hybrid models of CNN and Transformer have achieved great success in medical image segmentation, as well as in polyp segmentation tasks [15, 22]. However, polyp varies with different shapes, sizes, and morphology, thus it is always a challenging topic to achieve accurate polyp segmentation due to the characteristics of polyp images. In this work, we propose a

Table 2 Cross-evaluation of mDice scores for various segmentation methods across five polyp segmentation datasets: Kvasir, CVC-ClinicDB, CVC-ColonDB, ETIS, and CVC-300. The model was trained on Kvasir-SEG and CVC-ClinicDB datasets with 1450 images, and tested on the others. The bolded values represent the highest mDice scores for each dataset

Method	Kvasir	CVC-ClinicDB	CVC-ColonDB	ETIS	CVC-300
U-Net [8]	0.7615	0.8061	0.4734	0.5468	0.5969
U-Net++ [13]	0.8106	0.8769	0.4950	0.4762	0.6485
ResU-Net++ [9]	0.7581	0.8603	0.4007	0.3701	0.5027
HarDNet [36]	0.8892	0.9054	0.6139	0.8228	0.8793
U^2-Net [11]	0.8819	0.9204	0.6776	0.5543	0.7551
M^2SNet [12]	0.9045	0.9059	0.6042	0.8111	0.8479
TransUNet [15]	0.9187	0.9259	0.7382	0.8246	0.8692
UTNet [37]	0.8844	0.9263	0.6341	0.5458	0.8076
PGCF [7]	0.9216	0.9362	0.7687	0.8249	0.8742
CoinNet [38]	0.8861	0.8852	0.6207	0.5782	0.7561
DCATNet	**0.9266**	**0.9465**	**0.7872**	**0.8511**	**0.9064**

Table 3 Ablation study results for DCATNet on two benchmark polyp segmentation datasets: Kvasir-SEG and CVC-ClinicDB. The table evaluates the contribution of each proposed module by incrementally adding them to the baseline model. The baseline represents a simplified version of DCATNet without the specialized modules. The performance is measured using mDice and mIoU metrics, where higher values indicate better segmentation accuracy. The full DCATNet model, which integrates all modules, achieves the best performance on both datasets. "-DCN" means GAM with standard convolutional operations

Methods	Kvasir-SEG		CVC-ClinicDB	
	mDice	mIoU	mDice	mIoU
Baseline	0.9227	0.8566	0.9257	0.8617
Baseline + MSFE	0.9266	0.8633	0.9393	0.8855
Baseline + CAG	0.9285	0.8666	0.9369	0.8813
Baseline + GAM	0.9318	0.8724	0.9378	0.8829
Baseline + GAM-DCN	0.9217	0.8548	0.9294	0.8681
DCATNet	**0.9351**	**0.8781**	**0.9444**	**0.8948**

novel network architecture, called DCATNet. The proposed model incorporates Geometry Attention Modules to better capture shapes and boundary information. The Contextual Attention Gate (CAG) modules fuse features from the encoder and decoder with attention mechanism to further reduce the semantic gap. The MSEF is used to extract and fuse multi-scale features, which is based on the U-Block [11]. We validate the effectiveness of the proposed model on five public datasets with different scenarios, achieving better performance than most state-of-the-art medical image segmentation networks. The segmentation results across all the methods are visualized in Fig. 5. Additionally, we conduct ablation experiments to verify and explain the effectiveness of the proposed modules and network.

Table 4 Comparison of model size and inference time. Param is measured in million (M), floating point operations (FLOPs) are measured in Giga (G), and the inference time is measured by frames per second (FPS)

Method	Speed(FPS)	Param(M)	FLOPs(G)
U-Net [8]	502.94	11.32	14.06
U-Net++ [13]	190.18	13.27	30.29
ResUNet++ [9]	107.56	4.06	12.11
HarDNet [36]	63.05	33.34	4.62
U^2-Net [11]	53.13	44.02	28.87
M^2SNet [12]	31.95	29.74	6.91
TransUNet [15]	42.68	105.28	24.68
UTNet [37]	42.08	10.01	13.22
PGCF [7]	39.10	27.84	4.34
CoinNet [38]	23.12	44.66	36.1
DCATNet	42.13	108.39	28.91

The combination of CNN and Transformer can both extract local and global information, resulting in better segmentation performance, such as TransUNet [15] and Polyp-Pvt [16]. However, TransUNet is not proficient in capturing multi-scale features and modeling geometric features, which are very important for polyp segmentation tasks. Unlike TransUNet, the Polyp-Pvt uses the Pyramid Transformers as encoder, it integrates three submodules CFM, CIM, and SAM. The design of Polyp-Pvt is not well-suited for multi-scale feature extraction and fusion. To address these limitations, we introduce three key modules, including GAM, CAG, and MSFE. Our ablation results showed that the proposed module achieved better accuracy.

We conduct two individual experiments on Kvasir-SEG and CVC-ClinicDB, as well as cross-validation to verify the performance of the proposed model. The results show that DCATNet achieves significant improvements over previous SOTA methods across different datasets. The results of cross validation also show the strong robustness and generalization of our proposed model. Figures 6 and 7 show the dice score distribution of the test dataset by different methods. From that, we can also observe that the predictions of our method have a very small interval, which indicates the stability of our model.

To further validate the effectiveness of each module, we conduct an ablation study. We also perform the two-sample t-test on each component compared with the baseline model, all the results are listed in Table 3. From that, we can clearly observe the contribution of each component. The inclusion of GAM, CAG, and MSFE modules significantly improved the model's performance compared

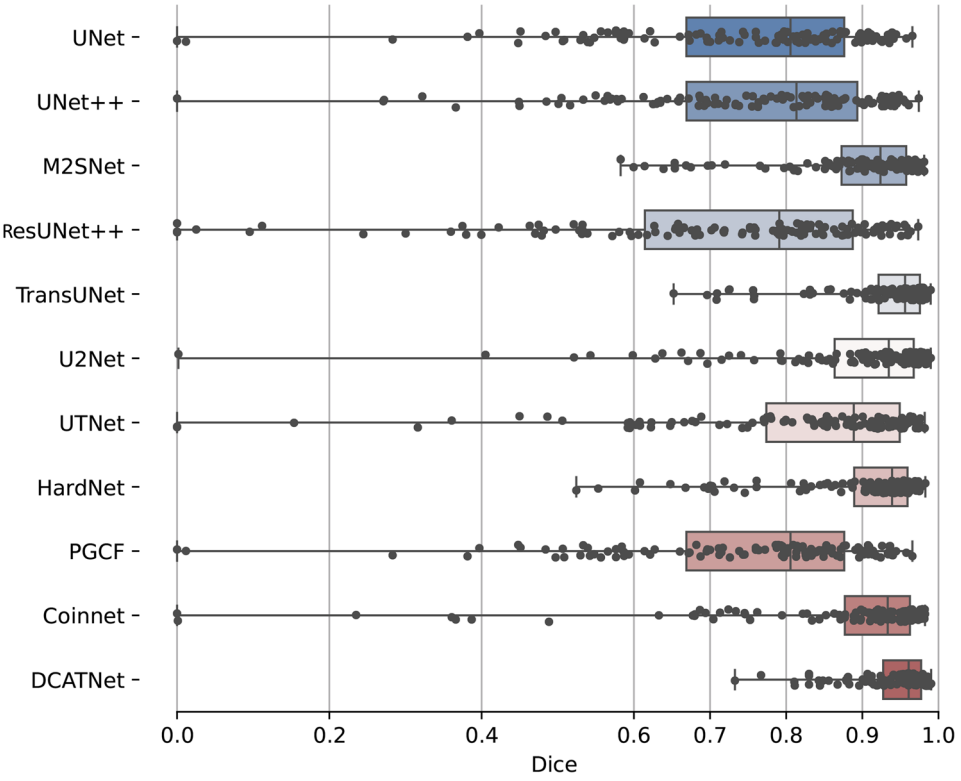

Fig. 6 Predicted Dice score distribution for different models on Kvasir-SEG dataset. The black dots represent the actual predicted dice score for each test image

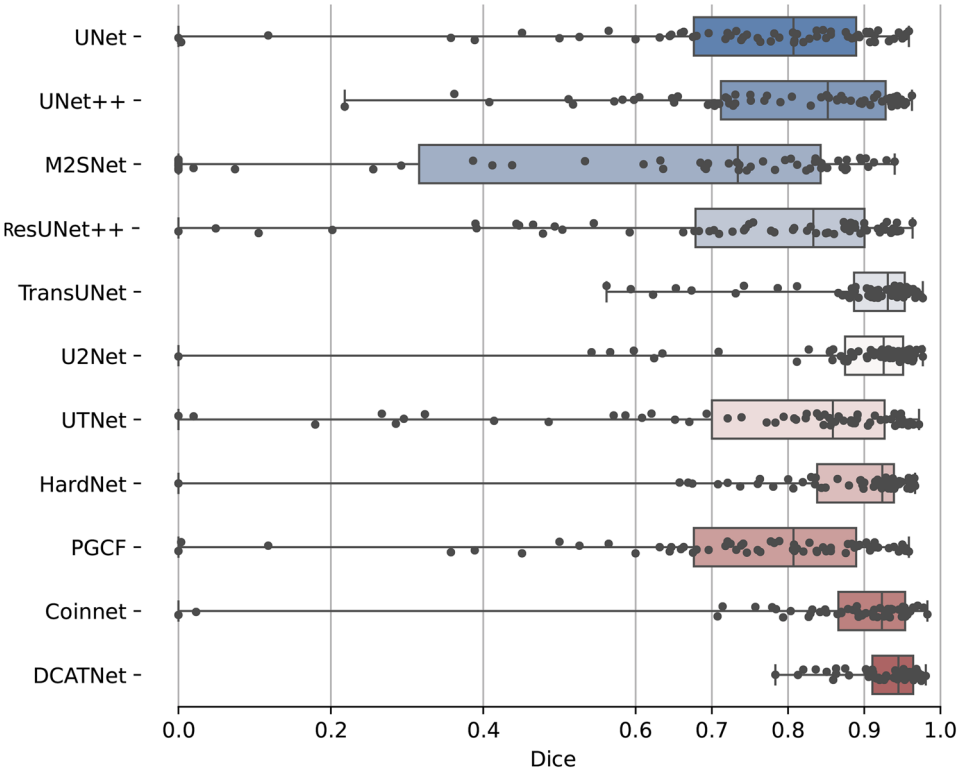

Fig. 7 Predicted Dice score distribution for different models on CVC-ClinicDB dataset. The black dots represent the actual predicted dice score for each test image

to the baseline. Furthermore, a small ablation study is conducted to verify the effectiveness of the deformable convolutions. This highlights the importance of adaptive feature extraction, context-aware attention mechanisms, and multi-scale feature extraction and integration in polyp image segmentation tasks.

There are a few limitations to our approach. Firstly, due to the 12-layer Transformer and deformable convolution operation, the complexity of the DCATNet model increases computational overhead and memory usage. This can be a hindrance in real-time applications or environments with limited computational resources. Secondly lacking multi-center medical data may affect the effectiveness and generalization of the model. In future work, we plan to gather and label images and videos from several institutions to meet the actual needs of the clinical environment. This would provide a more robust assessment of its generalizability and applicability. Additionally, we plan to investigate the impact of hyperparameter optimization on DCATNet's performance and explore strategies for further improving its accuracy and efficiency. With continued refinement and validation, it has the potential to make substantial contributions to clinical practice, improving early detection and treatment of colorectal polyps.

Conclusion

In this paper, we propose a novel approach, DCATNet, for polyp segmentation. The architecture integrates the Geometry Attention Module (GAM), Contextual Attention Gate (CAG) modules, and Multi-scale Feature Extraction (MSFE) module achieving a significant performance improvement and outperforming state-of-the-art models on the Kvasir-SEG and CVC-ClinicDB datasets. The combination of GAM, CAG, and MSFE decoders enhances feature representation and fusion, resulting in precise and reliable segmentation outcomes. These findings highlight the effectiveness of the proposed three modules in advancing medical image analysis.

Acknowledgements
Not applicable.

Author contributions
ZW, TL, ML, JJ, and XL designed and conceived the study. ZW and ML designed and conducted the experiments and drafted the manuscript. TL acquitted the data. ZW, JJ, and ML analyzed the results. TL, JJ, and XL reviewed the manuscript and provided critical feedback. All authors read and approved the final manuscript.

Funding
Not applicable.

Data availability
All datasets used in this paper are publicly available. The CVC-ClinicDB dataset is available at https://doi.org/10.1016/j.compmedimag.2015.02.007. The CVC-ColonDB dataset is obtained from https://doi.org/10.1109/TMI.2014.2314959. The ETIS-LaribPolypDB dataset is provided by https://doi.org/10.1007/s11548-013-0926-3. The Kvasir-SEG dataset can be accessed publicly at https://doi.org/10.1007/978-3-030-37734-2_37. The CVC-300 dataset can be accessed publicly at https://doi.org/10.1155/2017/4037190.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interest.

Received: 4 December 2024 / Accepted: 3 April 2025
Published online: 14 April 2025

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA A Cancer J Clinicians. 2021;71:209–49.
2. Leufkens A, van Oijen M, Vleggaar F, Siersema P.Factors influencing the miss rate of polyps in a back-to-back colonoscopy study. Endoscopy. 2012;44(5):470–75.
3. Van Rijn JC, Reitsma JB, Stoker J, Bossuyt PM, Van Deventer SJ, Dekker E.Polyp miss rate determined by tandem colonoscopy: a systematic review. Am J Gastroenterol. 2006;101(2):343.
4. Anderson R, Burr NE, Valori R.Causes of post-colonoscopy colorectal cancers based on world endoscopy organization system of analysis. Gastroenterology. 2020;158(5):1287–99.
5. Schottinger JE, Jensen CD, Ghai NR, Chubak J, Lee JK, Kamineni A, et al. Association of physician adenoma detection rates with postcolonoscopy colorectal cancer. Jama. 2022;327:2114–22.
6. Corley DA, Jensen CD, Marks AR, Zhao WK, Lee JK, Doubeni CA, et al. Adenoma detection rate and risk of colorectal cancer and death. N Engl J Med. 2014;370:1298–306.
7. Ji Z, Qian H, Ma X. Progressive Group Convolution Fusion network for colon polyp segmentation. Biomed Signal Process Control. 2024;96:106586.
8. Ronneberger O, Fischer P, Brox T U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer; 2015. p. 234–41.
9. Jha D, Smedsrud PH, Johansen D, de Lange T, Johansen HD, Halvorsen P, et al. A comprehensive study on colorectal polyp segmentation with ResUNet++, conditional random field and test-time augmentation. IEEE J Biomed Health Inf. 2021;25:2029–40.
10. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, et al. Attention u-net: learning where to look for the pancreas. 2018. arXiv preprint arXiv:180403999.
11. Qin X, Zhang Z, Huang C, Dehghan M, Zaiane OR, Jagersand M. U2-Net: going deeper with nested U-structure for salient object detection. Pattern Recogn. 2020;106:107404.
12. Zhao X, Jia H, Pang Y, Lv L, Tian F, Zhang L, et al. M2SNet: multi-scale in multi-scale subtraction network for medical image segmentation. arXiv preprint arXiv:230310894. 2023.
13. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J. 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. p. 3–11.
14. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16×16 words: transformers for image recognition at scale. 2020. arXiv preprint arXiv:201011929.
15. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al. TransUNet: transformers make strong encoders for medical image segmentation. CoRR. 2021. abs/2102.04306.
16. Dong B, Wang W, Fan DP, Li J, Fu H, Shao L. Polyp-pvt: polyp segmentation with pyramid vision transformers. 2021. arXiv preprint arXiv:210806932.
17. Zhang Y, Liu H, Hu Q Transfuse: fusing transformers and cnns for medical image segmentation. arXiv preprint arXiv:210208005. 2021.

18. Lin A, Chen B, Xu J, Zhang Z, Lu G, Zhang D Ds-transunet: dual swin transformer u-net for medical image segmentation. IEEE Transactions on Instrumentation and Measurement. 2022.
19. Duc NT, Oanh NT, Thuy NT, Triet TM, Dinh VS. Colonformer: an efficient transformer based method for colon polyp segmentation. IEEE Access. 2022;10:80575–86.
20. Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. SegFormer: simple and efficient design for semantic segmentation with transformers. In: Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Vaughan JW, editors. Advances in Neural Information Processing Systems. vol. 34. Curran Associates, Inc.; 2021. p. 12077–90. https://proceedings.neurips.cc/paper_files/paper/2021/file/64f1f27bf1b4ec22924fd0acb550c235-Paper.pdf
21. Shi W, Xu J, Gao P Ssformer: a lightweight transformer for semantic segmentation. In: 2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP). IEEE; 2022. p. 1–5.
22. Wang Z, Liu Z, Yu J, Gao Y, Liu M. Multi-scale nested UNet with transformer for colorectal polyp segmentation. J Appl Clin Med Phys. 2024;25(6):e14351. https://doi.org/10.1002/acm2.14351
23. Lai H, Luo Y, Zhang G, Shen X, Li B, Lu J.Toward accurate polyp segmentation with cascade boundary-guided attention. Visual Comput. 2023;39(4):1453–69.
24. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, et al. Swin-unet: unet-like pure transformer for medical image segmentation. In: European conference on computer vision. Springer; 2022. p. 205–18.
25. Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, et al. Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019. p. 3146–54.
26. Fan DP, Ji GP, Zhou T, Chen G, Fu H, Shen J, et al. Pranet: parallel reverse attention network for polyp segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer; 2020. p. 263–73.
27. Liu G, Yao S, Liu D, Chang B, Chen Z, Wang J, et al. CAFE-Net: cross-attention and feature exploration network for polyp segmentation. Expert Syst Appl. 2024;238:121754.
28. Wei X, Ye F, Wan H, Xu J, Min W. TANet: triple attention network for medical image segmentation. Biomed Signal Process Control. 2023;82:104608.
29. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–78.
30. Jha D, Smedsrud PH, Riegler MA, Halvorsen P, Lange T, Johansen D, et al. Kvasir-seg: a segmented polyp dataset. In: International Conference on Multimedia Modeling. Springer; 2020. p. 451–62.
31. Mamonov AV, Figueiredo IN, Figueiredo PN, Richard Tsai YH. Automated polyp detection in colon capsule endoscopy. IEEE Trans Med Imaging. 2014;33(7):1488–502. https://doi.org/10.1109/TMI.2014.2314959.
32. Bernal J, Sánchez FJ, Fernández-Esparrach G, Gil D, Rodrguez C, Vilariño F. WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. Comput Med Imaging Graphics. 2015;43:99–111.
33. Silva J, Histace A, Romain O, Dray X, Granado B. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. Int J Comput Assisted Radiol Surg. 2014;9(2):283–93. https://doi.org/10.1007/s11548-013-0926-3
34. Vázquez D, Bernal J, Sánchez FJ, Fernández-Esparrach G, López AM, Romero A, et al. A benchmark for endoluminal scene segmentation of colonoscopy images. J Healthcare Eng. 2017;2017:4037190.
35. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: an imperative style, high-performance deep learning library. Adv Neural Inf Process Syst. 2019;32:8026–37.
36. Huang CH, Wu HY, Lin YL. HarDNet-MSEG: a simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 FPS. 2021. arXiv preprint arXiv:210107172.
37. Gao Y, Zhou M, Metaxas DN. UTNet: a hybrid transformer architecture for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. Springer; 2021. p. 61–71.
38. Jain S, Atale R, Gupta A, Mishra U, Seal A, Ojha A, et al. Coinnet: a convolution-involution network with a novel statistical attention for automatic polyp segmentation. IEEE Trans Med Imaging. 2023;42:3987–4000.

Publisher's Note