scientific reports

Check for updates

OPEN

Hierarchical contextual information aggregation for polyp segmentation

Lin Li[1], Haichen Yang[2], Jialing Zhang[3], Cuijuan Zheng[1], Xiaoyu Chen[4] & Xue He[1]✉

Accurate segmentation of polyp tissues in colonoscopic images is crucial for early colorectal cancer detection. Existing CNN-based approaches effectively capture local dependencies but struggle with long-range relations, while transformer-based methods excel in global context modeling yet often overlook fine contextual details. Hybrid CNN–transformer models attempt to combine both, but typically overfit to convolutional features, weakening attention mechanisms. To address these limitations, we propose a Hierarchical Contextual Information Aggregation Network (HCIA) for polyp segmentation. HCIA introduces an Interconnected Attention Module (IAM) that applies global attention to single-level features, enabling comprehensive cross-hierarchy information exchange. In parallel, a Hierarchical Aggregation Module (HAM) fuses adjacent feature levels to enhance local contextual representation. This dual refinement allows HCIA to jointly capture global and local dependencies, yielding more precise tissue boundaries. Extensive experiments across multiple polyp segmentation benchmarks demonstrate that HCIA achieves superior generalization and state-of-the-art accuracy, highlighting its potential for clinical applications.

Keywords Polyp segmentation, Deep learning, Transformer, Generalization

Colorectal cancer (CRC) remains a leading cause of cancer-related mortality worldwide[1]. Since most CRC cases originate from adenomatous polyps, early detection and removal via colonoscopy are critical for prevention. However, manual segmentation of polyps is labor-intensive and prone to human error, with missed detection rates ranging from 17% to 28%[2,3]. Consequently, there is an urgent demand for automated, high-precision computer vision systems to assist clinicians in polyp segmentation.

In recent years, deep learning has revolutionized this field. CNN-based methods, such as U-Net[4] and its variants[5,6], have demonstrated strong capabilities in modeling local features but often struggle to capture long-range dependencies due to limited receptive fields. Conversely, Transformer-based models[7] excel at global context modeling but may lose fine-grained local details and suffer from high computational complexity. To mitigate these issues, recent hybrid approaches (e.g., TransNetR[8]) attempt to combine convolutional layers with self-attention mechanisms. Despite their progress, existing hybrid models still face significant limitations: 1) the intrinsic discriminative capacity of convolutional features is not fully exploited within the hybrid feed-forward paths; and 2) they often exhibit diffuse attention distribution, leading to insufficient boundary delineation and an inability to effectively balance global coherence with local precision (Fig. 1).

To address these specific limitations, we propose the Hierarchical Contextual Information Aggregation Network (HCIA). Unlike existing methods, HCIA is explicitly designed to establish stable long-range dependencies while preserving local structural details through two novel components. First, the **Interconnected Attention Module (IAM)** captures global dependencies across all hierarchical levels using a shared memory mechanism with linear complexity, ensuring comprehensive supervision. Second, the **Hierarchical Aggregation Module (HAM)** facilitates the dynamic integration of multi-scale features from adjacent layers, effectively suppressing background noise and sharpening polyp boundaries. Extensive experiments demonstrate that HCIA achieves superior accuracy and generalization compared to state-of-the-art methods.

The main contributions of this paper are summarized as follows:

[1]Department of Anesthesia Surgery, The Affiliated Huaian No.1 People's Hospital of Nanjing Medical University, No.1, Huai'an 223300, China. [2]Department of Emergency, The Affiliated Huaian Hospital of Xuzhou Medical University and Huaian Second Peoples Hospital, No.1, Huai'an 223300, China. [3]Department of Gastroenterology, The Affiliated Huaian No.1 People's Hospital of Nanjing Medical University, Huai'an 223300, China. [4]Department of Radiology, The Affiliated Huaian Hospital of Xuzhou Medical University and Huaian Second Peoples Hospital, Huai'an 223300, China. ✉email: 18915110368@163.com

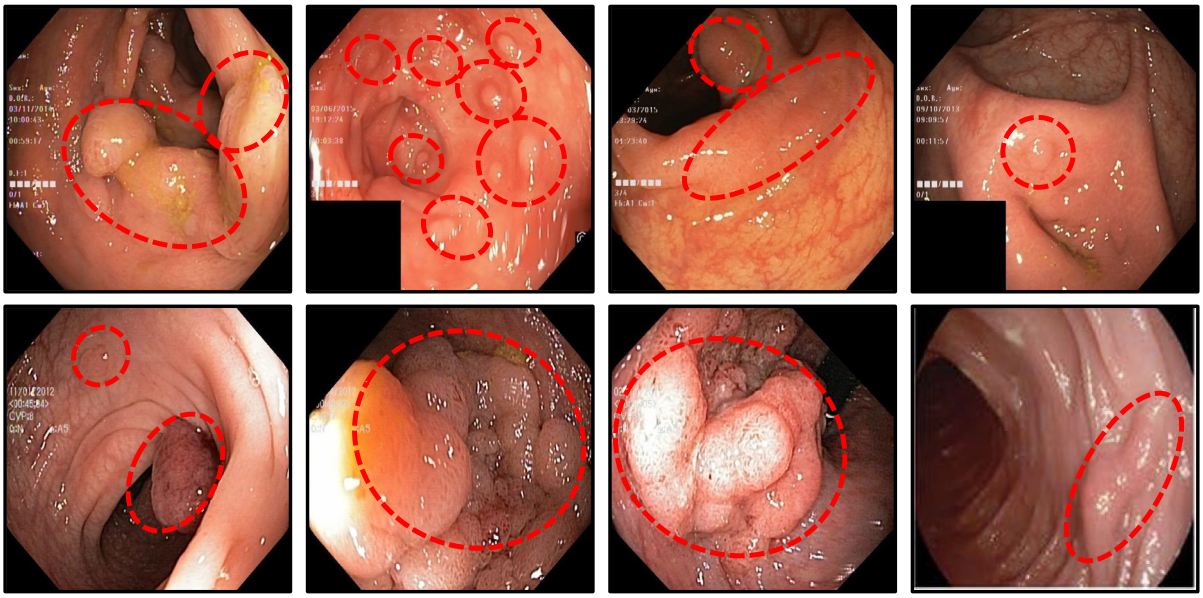

Fig. 1. Illustration of the typical colorectal polyps.

- We propose the HCIA network, a novel architecture that efficiently aggregates hierarchical contextual information for precise polyp segmentation.
- We introduce the Hierarchical Aggregation Module (HAM) to integrate multi-scale features from adjacent branches, enhancing boundary discrimination and robustness against variable polyp sizes.
- We design the Interconnected Attention Module (IAM) to establish global dependencies across hierarchical levels with linear complexity, enabling effective global supervision.
- Comprehensive evaluations on multiple benchmarks demonstrate that HCIA achieves state-of-the-art performance and exhibits strong generalization capabilities.

Related work

CNN-based segmentation methods

Polyp segmentation classifies pixels from colonoscopy images into polyp tissue categories, generating an accurate tissue mask for subsequent clinical analysis. Initially, CNNs dominated this field for their proficiency in modeling local contextual information, thereby becoming the go-to architecture for such tasks for years. Pioneering this domain, Brandao et al.[9] were the first to leverage fully convolutional networks specifically for this purpose. The introduction of the UNet model by Ronneberger et al.[4], with its innovative encoder-decoder and skip connection strategy, marked a significant leap, facilitating dense and high-resolution predictions. Building on this foundational work, a plethora of UNet-inspired architectures emerged, with notable improvements such as the nested dense skip connections found in Zhou et al.'s UNet++[5], and the pyramid-style feature coding for multi-scale representations in Jha et al.'s ResUNet++[6]. Departing from traditional U-shaped blueprints, alternate designs like Fan et al.'s PraNet[10] capitalized on reverse attention to enhance edge delineation of polyp tissues, and Tomar et al.'s DDANet[11] employed a dual-decoder mechanism to furnish additional attention-guided maps. Zhao et al.[12] introduced the TACT network, which employs the FAPS module for precise segmentation of polyp boundaries and integrates high-level and low-level features through the MSFA module. Huang et al.[13] proposed MGF-Net, which refines polyp edge details with enhanced accuracy via multi-channel grouping fusion. Liu et al.[14] developed the multi-cascade network MCA-Net, focusing on issues related to variations in polyp shape, size, and texture. Du et al.[15] presented UM-Net, which mitigates the impact of polyp tissue color using a color transfer operation. Zhu et al.[16] introduced Polyp-Mamba, which employs discrete cosine transform to analyze features from multiple spectral perspectives. Nevertheless, despite the progression in CNN-derived approaches, their intrinsic inability to adeptly model long-range dependencies has surfaced as an intrinsic bottleneck, presenting a barrier to the evolution of the segmentation proficiency.

Transformer-based segmentation methods

The transformative prowess of transformers in capturing long-range dependencies has seen them gain traction in various fields, including polyp segmentation. Recent contributions, such as by Wang et al.[17], devised the SSFormer employing a pyramid transformer backbone for elevated segmentation accuracy. Meanwhile, Duc et al.[18] constructed a nimble model dubbed ColonFormer that serves as a segmentation baseline. Despite these innovations, transformer-based techniques are somewhat hamstrung by their native self-attention mechanisms, which do not adequately account for context-rich dependencies.

In an attempt to bridge this gap, some researchers have embarked on crafting hybrid networks that integrate convolutional computations into the transformer framework or merge self-attention and convolutional layers.

Such models strive to capture both long-range and close-knit dependencies concurrently. Zhang et al.[19] put forward TransFuse, a model which synchronously harnesses parallel CNN and transformer encoders to learn both global and local relationships. Similarly, the TransUNet, as showcased by Chen et al.[20], chains transformers to CNNs for multi-scale feature synthesis, and Dong et al.[21] introduced the poly-PVT that employs the Pyramid Transformer to reach into hierarchically structured features, enhanced by multiple decoders for pixel-wise segmentation. He et al.[22] proposed CTHP, which utilizes unidirectional attention to comprehensively model both global and local information. Liu et al.[23] introduced CAFÉ-Net, designed to maximize the utilization of fine-grained information by reconstructing missing data while preserving low-level features. Xiao et al.[24] presented CTNet, which enhances segmentation accuracy by leveraging multi-scale information and high-resolution features. Wang et al.[25] developed WBANet, focusing on modeling multi-scale edge information by extracting the slope of the polyp tissue edges. Notably, these hybrid designs have demonstrated superior performance; however, two principal challenges persist: 1) The intrinsic discriminative capacity of convolutional features integrated within the feed-forward layers is not fully realized, and 2) a propensity for hybrid models to overly conform to convolutional features, potentially resulting in diffuse attention that undermines the overall efficacy of the model.

In contrast to these methods, HCIA does not establish long-range dependencies and local context solely through Transformer or convolution operations; instead, it achieves this through the interplay of IAM and HAM. The IAM refines the feature representation of the current layer by calculating attention over the hierarchical features at each level, facilitating the exchange of information across different scales via a shared attention memory. Meanwhile, the HAM is responsible for connecting adjacent layers and integrating their features, thereby obtaining a multi-scale fused perspective. This design enables HCIA to establish stable long-range dependencies and local information while effectively mitigating the drawbacks associated with hybrid models.

Attention module in polyp segmentation

Due to the need for fine-grained features in polyp segmentation, numerous methods have employed various attention modules to enhance retrieval performance. Fan et al.[10] utilized reverse attention to further optimize the extraction of details at the edges of polyp tissues. Zhang et al.[26] adopted cross-semantic attention to calibrate low-dimensional semantic information within the encoder. Liu et al.[27] implemented convolution-based attention to focus on locally significant information. He et al.[22] employed height-direction and weight-direction attention to model local contextual features. Liu et al.[23] utilized a cross-attention decoder to protect low-level features while recovering fine-grained characteristics. Wang et al.[25] designed a hierarchical attention fusion mechanism to guide the model's focus towards critical regions.

However, most existing attention modules typically prioritize local subtle features or suffer from high computational costs. To explicitly clarify the novelty of our method, we summarize the key distinctions between the proposed IAM/HAM and closely related attention mechanisms as follows:

- **Global Supervision** *vs* **. Local Refinement:** Unlike Reverse Attention[10] or Convolution-based Attention[27], which are limited to refining local edges or specific regions, our IAM leverages a globally shared memory. This allows for information exchange across all network branches and layers, forming a coherent global supervision framework rather than isolated local enhancements.
- **Linear** *vs* **. Quadratic Complexity:** While Cross-Semantic Attention[26] and standard self-attention mechanisms generally exhibit quadratic complexity $O(N^2)$, imposing a heavy computational burden, IAM is designed with linear complexity. This significantly improves computational efficiency without sacrificing the ability to model global dependencies.
- **Inter-layer Interaction** *vs* **. Single-layer Focus:**In contrast to methods that apply attention independently within specific layers (e.g[22].,), our HAM explicitly connects adjacent layers. This integration fuses multi-scale features dynamically, ensuring that both high-level semantic guidance and low-level structural details are preserved effectively.

Proposed method

Encoder

To effectively capture multi-scale hierarchical features conducive to polyp segmentation, our framework adopts PVTv2[28], initially pretrained on the ImageNet[29], to serve as the backbone encoder, as shown in Fig. 2 (a). PVTv2 distinguishes itself from traditional vision transformers by incorporating self-attention blocks married with strided convolutions. This strategic amalgamation allows for the formation of long-range spatial dependencies across a descending cascade of feature resolutions—a design tuned for dense prediction tasks that simultaneously seeks to curtail the computational expense. The outcome is a pyramidal architecture that yields a multi-tier suite of features spanning various scales. More specifically, the encoder provides a quartet of hierarchical feature levels, designated as $f_i \in \mathbb{R}^{\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i}$ for $i \in \{1, 2, 3, 4\}$. Each subsequent level F_i not only diminishes in spatial dimensions, enabling a larger receptive field, but also exhibits escalated feature dimensionality, reflecting the increased abstraction of deeper levels. For decoding purposes, these features undergo channel-wise refinement via 1×1 convolutions, preparing a tailored input for the decoder branch of our architecture.

Hierarchical aggregation module

The hierarchical aggregation module (HAM) plays a pivotal role in assimilating multi-scale hierarchical features across adjacent branches, an essential process for the effective attenuation of background interference and the enhancement of salient foreground regions. This functionality is crucial when dealing with polyps across a

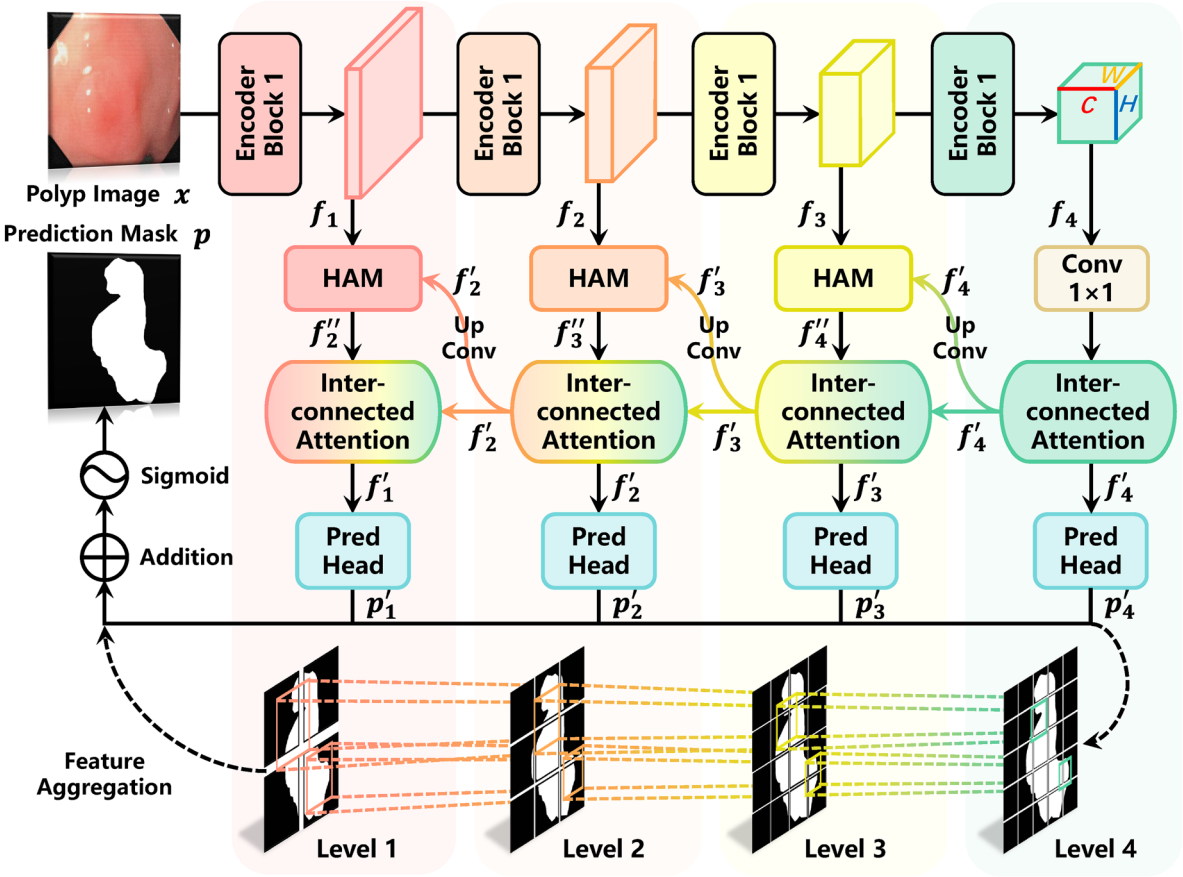

Fig. 2. Illustration of the proposed Hierarchical Contextual Information Aggregation Network (HCIA), which consists of four hierarchical branches. The output local prediction maps $p_1\prime$, $p_2\prime$, $p_3\prime$, and $p_4\prime$ are aggregated as the final prediction p. 'HAM' denotes the Hierarchical Aggregation Module (HAM). 'Inter Connected Attention' denotes the Interconnected Attention Module (IAM). 'Pred Head' represents the Prediction Head. 'UpConv' is an upsampling convolutional layer.

spectrum of sizes and for the discernment of clear and reliable boundaries—both vital attributes for bolstering the generalizability of the segmentation model. Drawing from the attention mechanisms presented by Oktay et al.[30], our HAM implementation utilizes a grid-attention framework.

As depicted in Fig. 2, the operations within the HAM at the $i-th$ ($i \in \{1,2,3\}$) hierarchical level involve harmonizing intermediate features f'_{i+1} from the antecedent lower-level branch together with the concurrent level's features f_i. This integration enables our model to concurrently optimize for local feature refinement and global context, which is instrumental for achieving high-fidelity polyp segmentation. This process can be formulated as:

$$f'_{i+1} = \mathrm{UpConv}(f'_{i+1}), \tag{1}$$

$$f_i = \mathrm{GN}(\mathrm{Conv}_{1\times1}(f_i)), \quad f'_{i+1} = \mathrm{GN}(\mathrm{Conv}_{1\times1}(f'_{i+1})), \tag{2}$$

$$A^{HAM}_{i,i+1} = \mathrm{GN}(\mathrm{Conv}_{1\times1}(\sigma(f_i + f'_{i+1}))), \tag{3}$$

$$f''_i = \left[\tau(A^{HAM}_{i,i+1}) \odot f_i, f'_{i+1}\right], \tag{4}$$

where $\mathrm{UpConv}(\cdot) = \sigma(\mathrm{GN}(\mathrm{Conv}_{3\times3}(\mathrm{Up}_{2\times}(\cdot))))$,$\mathrm{Up}_{2\times}$ denotes upsampling operation. $\mathrm{Conv}_{3\times3}$ is a convolution layer with a 3×3 kernel while $\mathrm{Conv}_{1\times1}$ is that with a 1×1 kernel. GN denotes a group normalization. $A^{HAM}_{i,i+1}$ denotes an affinity matrix. σ is the SiLU function. τ is the Sigmoid function. $\odot$ denotes matrix multiplication and [,] represents the concatenation operation.

Interconnected attention module

The Interconnected attention module (IAM) is a cornerstone component of the proposed HCIA that fulfills a dual purpose. First, the IAM forges local contextual dependencies among the varying levels of hierarchical branches, effectively enriching the capacity of these features to distinguish relevant patterns within the data. Second, it leverages globally shared memories, designated as M_k and M_v, that are intricately woven into the attention

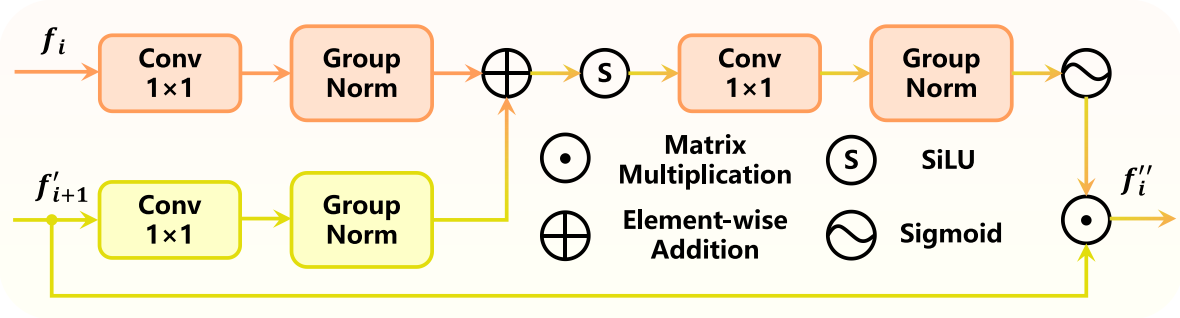

Fig. 3. Illustration of the structure of Hierarchical Aggregation Module (HAM).

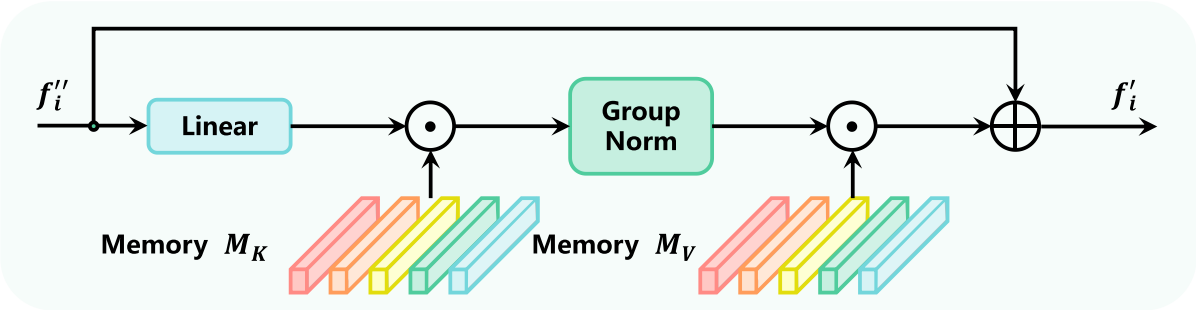

Fig. 4. Illustration of the mechanism of Interconnected Attention Module (IAM).

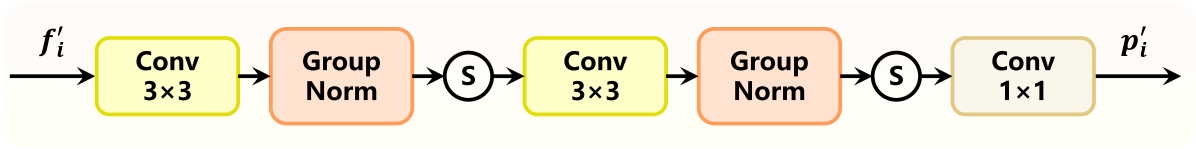

Fig. 5. Illustration of the structure of Prediction Head (PH).

computations of IAMs across all four hierarchical branches. This incorporation of global memories imparts a comprehensive supervisory element, orchestrating an integrated approach to information consolidation across the multiple scales of hierarchical features (Figs. 3, 4 and 5).

In our construction of IAMs, we employ an external attention mechanism as formulated by Guo et al.[31], which allows for an expansive attention reach beyond the confines of local associations typically seen in convolutional networks, as shown in Fig. 2 (c). The use of external attention not only provides a means to capture higher-order dependencies within the feature maps but also ensures that the attention mechanism benefits from a consistency in the oversight across all levels of feature hierarchy. This unified attentional guidance is intended to significantly bolster the accuracy in segmenting polyps of various scales and complexities, delivering superior model performance.This process can be formulated as:

$$A_i^{IAM} = (W_i^{IAM})^\top f_i'' \odot M_k, \tag{5}$$

$$f_i' = f_i'' + \mathrm{GN}(A_i^{IAM}) \odot M_v, \tag{6}$$

where W_i^{IAM} is a weight matrix of a learnable linear layer in the $i-th$ hierarchical branch. A_i^{IAM} denotes an affinity matrix. $\odot$ represents matrix multiplication. M_k and M_v are globally shared memories which are randomly initialized.

Indeed, in the Interconnected Attention Module (IAM) design within the Hierarchical Contextual Information Aggregation Network (HCIA), the substitution of the standard keys and values, typically computed from the input features f_i'', with the pre-established globally shared memories M_k and M_v is a strategic innovation. This replacement abandons the notoriously computation-intensive quadratic complexity of traditional self-attention schemes for a far more efficient linear complexity framework. This shift to linear complexity in the attention computation confers a substantial enhancement in the computational efficiency of the IAM. By circumventing

the demand for pairwise feature comparisons intrinsic to quadratic attention, the IAM can perform attention operations over extensive feature maps without incurring debilitating computational costs. This efficiency gain propels the Hierarchical Contextual Information Aggregation Network (HCIA) to a more feasible status compared to counterparts that are still anchored to the conventional quadratic computational paradigm, particularly in the context of large input dimensionality or extensive datasets—commonplace in medical image analysis.

Prediction head

The prediction head is responsible for generating local prediction maps from the outputs of each hierarchical branch in the architecture, as shown in Fig. 2 (e). For the $i-th$ hierarchical layer, the produced local prediction map is represented as p_i'. To enhance the robustness and comprehensiveness of the feature discrimination, we aggregate the local prediction maps $p_1\prime, p_2\prime, p_3\prime$, and $p_4\prime$ emanating from all four hierarchical branches. This summation harnesses the combined strengths of each level's feature representation to yield a more holistic and refined feature profile. The collective map is then further refined by passing it through a Sigmoid activation function. This process can be formulated as:

$$p_i' = \mathrm{Conv}_{1\times1}(\mathrm{PRE}_2(\mathrm{PRE}_1(f_i'))), \tag{7}$$

$$p = \tau(\sum_{i=1}^{4} p_i'), \tag{8}$$

where $\mathrm{PRE}(\cdot) = \sigma(\mathrm{GN}(\mathrm{Conv}_{3\times3}(\cdot)))$ is a prediction module, which consists of a convolution layer with a kernel of 3×3, a group normalization layer, and a SiLU activation function. τ is the Sigmoid function. p denotes the overall prediction map.

Experiments

Datasets and metrics

Kvasir-SEG. The Kvasir-SEG dataset[32] is a publicly available collection featuring 1000 polyp images, complete with expertly annotated segmentation masks verified by seasoned gastroenterologists. These images pose a substantial challenge due to their diversity in size—ranging from dimensions of 332×487 to 1920×1072—and the variety in the appearance of the polyps themselves, which differ markedly in size, shape, and texture.

CVC-ClinicDB. The CVC-ClinicDB dataset[33] is a widely accessible set of 612 colonoscopy images derived from 31 video sequences, standardized to a resolution of 384×288. Accompanying each image is a meticulously annotated, pixel-accurate segmentation mask, validated by medical experts to ensure reliability.

CVC-ColonDB. The CVC-ColonDB dataset[34] comprises a total of 380 annotated images extracted from 15 video sequences, with each image having a resolution of 574×500. Each image has been validated by medical experts to exclude similar images, ensuring that the dataset represents content from different perspectives.

Evaluation metrics. Our model's efficacy is gauged using four established metrics[10,35], each serving as a standard for performance evaluation in polyp segmentation: mean Dice coefficient (*mDice*)[36], mean Intersection over Union (*mIoU*), Structure-measure (S_α)[37], and Enhanced-alignment measure (mE_ξ)[38]. The *mDice* and *mIoU* offer quantifiable insights into the region-centric similarity between predicted and ground truth masks. S_α evaluates the structural integrity of the predicted segmentation by aligning it with the object-attuned similarity. Finally, mE_ξ delivers a nuanced index of the segmentation's fidelity by assessing the model's predictive competency at both the global image level and the detailed pixel level. We conduct five independent tests on the model, and the average of the evaluation metrics from these five tests is taken as the final result to mitigate any potential variability or bias.

Implementation details

Our architecture is instantiated and investigated within the PyTorch version 1.11.0. For the training and assessment phases, we leverage the computational prowess of an NVIDIA RTX 3090 GPU outfitted with a substantial 24GB of VRAM. To standardize the input data, we resize all polyp images to a uniform dimension of 352×352 pixels. We deploy a suite of image augmentation strategies to enhance the generalizability of the model, including Gaussian blur, color jittering, horizontal and vertical flips. Additionally, affine transformations are applied to simulate common variations in positioning, such as translation, rotation, scaling, and shearing. The model optimization process harnesses the Adam[53]. We initiate training with a learning rate set at $1e-4$. To facilitate a streamlined training operation capable of accommodating sizable datasets, we adopt a batch size of 16 and employ mixed-precision training through NVIDIA's Apex library, which accelerates computation while reducing memory requirements.

Evaluation

The proposed HCIA in the comparison presented, exhibits exceptional effectiveness and generalizability against a selection of contemporary state-of-the-art (SOTA) models on the public datasets Kvasir-SEG, CVC-ClinicDB and CVC-ColonDB. For a comprehensive assessment, HCIA is benchmarked against two distinct groups of SOTA methodologies: 11 CNN-based models, including U-Net[4], UNet++[5], PraNet[10], DCRNet[40], MMFIL-Net[42], EFA-Net[45], SRaNet[47], BUNet[48], MADGNet[41], Polyp-Mamba[16], UM-Net[15] and APCNet[44], as well as 10 Transformer-derived models, namely Transfuse[19], Polyp-PVT[46], Polyp-LVT[43], SSFormer[49], FCBFormer[52], CAFE-Net[23], CTNet[24], DSHNet[50], CTHP[22] and MGCBFormer[51]. To ensure fair comparison, all open-source methods, including ours, are retrained with authors' publicly available source codes on a unified hardware setup

maintaining their default configurations. The assessment encompasses two dimensions: efficacy, as demonstrated on individual dataset testing subsets, and broader applicability through cross-dataset evaluations. Models are trained on specified splits from each dataset and then tested within their corresponding domains. For cross-dataset generalizability, models trained on the entirety of Kvasir-SEG are evaluated on CVC-ClinicDB (Kvasir $\rightarrow$ CVC), and vice versa.

Reflecting on the results noted in Table 1, HCIA presents a compelling performance profile, consistently outshining other methods on most metrics across the board for the datasets under study. Specifically, HCIA leads on Kvasir-SEG, achieving 94.2% on *mDice*, 89.4% on *mIoU*, 94.9% on S_α and 96.8% on mE_ξ, outmatching CNN-based method Polyp-Mamba[16] (2025) in terms of *mDice* and *mIoU* by 2.3% and 2.7%, while surpassing Transformer-based method DSHNet[50] in terms of *mDice* and *mIoU* by 1.7% and 1.3%. On CVC-ClinicDB and CVC-ColonDB, similarly, HCIA's performance expressed in nearly all metrics surpasses that of competing CNN-based methods and Transformer-based solutions, solidifying the contribution of the hierarchical and attentive components HAM and IAM to the overall performance.

Table 2 lays out the generalization results which underscore relatively excellent capacity of HCIA to maintain its stable performance across disparate datasets. In the Kvasir $\Rightarrow$ CVC setting, HCIA shows a striking advantage over FCBFormer-L in *mDice* and *mIoU* by 4% and 4.4%, while for CVC $\rightarrow$ Kvasir, HCIA surpasses FCBFormer-L in *mDice* and *mIoU* by 5.8% and 9%. This consistency across diverse training and testing scenarios underlines the robustness of the HCIA's design ethos in tackling the intricate task of polyp segmentation.

Ablation studies

The meticulous ablation study conducted on HCIA sheds light on the potency and pivotal roles of the constituent modules, primarily the Hierarchical Aggregation Module (HAM) and the Interconnected Attention Module (IAM), across Kvasir-SEG and CVC-ClinicDB. As delineated in Table 3 and Table 4, these experiments are bifurcated into incremental and comparative sets.

As shown in Table 3, the incremental experiments progressively enrich the baseline model with core components—the HAM and the IAM—examining their individual and combined influence on model performance. On the other hand, as shown in Table 4, the comparative experiments evaluate different attention mechanisms within the IAM framework by testing a series of derivative variants. Variant I^∇ substitutes the external attention with spatial attention (SA), $I^\triangle$ with channel attention (CA), $I^\square$ harnesses both SA and CA simultaneously, and $I^\diamond$ utilizes dedicated M_k and M_v memories for each IAM, shedding the globally shared property.

Methods	Kvasir-SEG				CVC-ClinicDB				CVC-ColonDB			
	mDice	*mIoU*	S_α	mE_ξ	*mDice*	*mIoU*	S_α	mE_ξ	*mDice*	*mIoU*	S_α	mE_ξ
U-Net[4]	0.815	0.742	0.854	0.877	0.818	0.751	0.883	0.907	0.550	0.465	0.720	0.724
UNet++[5]	0.817	0.740	0.858	0.882	0.792	0.727	0.873	0.890	0.498	0.411	0.685	0.718
MedSAM[39]	0.862	0.795	-	-	0.867	0.803	-	-	0.734	0.651	-	-
PraNet[10]	0.895	0.844	0.909	0.941	0.895	0.844	0.936	0.962	0.701	0.632	0.814	0.835
DCRNet[40]	0.889	0.825	0.911	0.935	0.893	0.840	0.933	0.965	0.723	0.650	0.817	0.862
MADGNet[41]	0.907	0.853	0.856	0.947	0.939	0.895	0.922	0.985	0.775	0.697	0.833	0.880
MMFIL-Net[42]	0.909	0.858	0.915	0.948	0.890	0.838	0.922	0.971	0.744	0.659	0.824	0.864
Polyp-LVT[43]	0.909	0.851	-	0.941	0.935	0.882	-	0.982	-	-	-	-
APCNet[44]	0.913	0.859	0.919	-	0.934	0.886	0.945	-	0.758	0.682	0.837	-
EFA-Net[45]	0.914	0.861	0.929	0.955	0.919	0.871	0.943	0.972	0.774	0.696	0.855	0.884
CTNet[24]	0.917	0.863	0.928	0.959	0.936	0.887	0.952	0.983	0.813	0.734	0.874	0.915
Polyp-PVT[46]	0.918	0.861	0.924	0.955	0.938	0.886	0.947	0.981	0.807	0.726	0.864	0.913
Polyp-Mamba[16]	0.919	0.867	**0.951**	0.968	0.941	0.896	**0.970**	0.987	0.791	0.713	0.846	0.902
Transfuse[19]	0.920	0.871	0.926	0.959	0.918	0.868	0.935	0.972	0.702	0.782	0.822	0.886
SRaNet[47]	0.921	0.870	0.932	-	0.926	0.875	0.950	-	0.814	0.734	**0.877**	-
BUNet[48]	0.923	0.873	0.926	0.965	0.935	0.890	0.948	0.989	0.813	0.731	0.861	0.911
SSFormer-S[49]	0.926	0.874	-	-	0.926	0.875	-	-	0.771	0.711	-	-
DSHNet[50]	0.929	0.881	0.936	0.965	0.942	0.896	0.954	0.987	0.815	0.733	0.870	0.915
UM-Net[15]	0.930	0.925	0.938	-	-	-	-	-	0.761	0.828	0.863	-
CAFE-Net[23]	0.933	0.889	0.939	0.967	0.943	0.899	0.957	0.986	0.820	0.740	0.874	0.914
MGCBFormer[51]	0.933	0.887	-	-	0.936	0.899	-	-	0.731	0.807	-	-
SSFormer-L[49]	0.935	0.890	-	-	0.944	0.899	-	-	0.804	0.722	-	-
FCBFormer[52]	0.938	0.890	0.945	0.940	0.946	0.902	0.952	0.944	0.777	0.696	0.841	0.893
CTHP[22]	0.939	0.891	-	-	0.947	0.902	-	-	-	-	-	-
HCIA(ours)	**0.942**	**0.894**	0.949	**0.968**	**0.948**	**0.904**	0.959	**0.990**	**0.823**	**0.830**	0.874	**0.920**

Table 1. Single-domain performance evaluation of HCIA compared to other SOTA models. Optimal results are highlighted in bold.

Methods	Kvasir → CVC		CVC → Kvasir	
	mDice	*mIoU*	*mDice*	*mIoU*
U-Net[4]	0.589	0.501	0.522	0.401
ResUNet[32]	0.333	0.236	0.328	0.217
ResUNet++[6]	0.560	0.470	0.302	0.204
Deeplabv3+[54]	0.650	0.538	0.674	0.532
PraNet[10]	0.784	0.707	0.791	0.703
MSRF-Net[55]	0.792	0.649	0.757	0.633
DCRNet[40]	0.724	0.630	0.727	0.631
Transfuse[19]	0.772	0.673	0.768	0.669
SSFormer-S[49]	0.796	0.722	0.779	0.697
SSFormer-L[49]	0.833	0.757	0.827	0.734
FCBFormer[52]	0.873	**0.803**	0.884	0.821
HCIA(ours)	**0.873**	0.801	**0.885**	**0.824**

Table 2. Cross-domain performance evaluation of HCIA compared to other SOTA models. Optimal results are highlighted in bold.

Methods	Kvasir-SEG				CVC-ClinicDB				Params	FLOPs
	mDice	*mIoU*	S_α	mE_ξ	*mDice*	*mIoU*	S_α	mE_ξ		
Base	0.905	0.854	0.915	0.949	0.915	0.864	0.926	0.966	20.5	10.4
Base+H	0.923	0.879	0.940	0.957	0.928	0.884	0.938	0.976	24.4	12.5
Base+I	0.928	0.883	0.939	0.958	0.936	0.879	0.943	0.975	21.4	11.0
Base+H+I	0.942	0.894	0.949	0.968	0.948	0.904	0.959	0.990	25.8	13.7

Table 3. The ablation studies of HCIA evaluated in single domain on Kvasir-SEG and CVC-ClinicDB. *Base* denotes the adopted baseline. *H* denotes HAM while *I* denotes IAM. Params denotes model parameters. FLOPs denotes floating point operations.

Methods	Kvasir-SEG				CVC-ClinicDB			
	mDice	*mIoU*	S_α	mE_ξ	*mDice*	*mIoU*	S_α	mE_ξ
Base+H+I$^\nabla$	0.930	0.886	0.940	0.959	0.939	0.892	0.946	0.979
Base+H+I$^\triangle$	0.931	0.885	0.938	0.957	0.938	0.890	0.943	0.976
Base+H+I$^\square$	0.935	0.889	0.941	0.959	0.941	0.896	0.955	0.982
Base+H+I$^\diamond$	0.936	0.890	0.945	0.962	0.943	0.898	0.985	0.981
Base+H+I	0.942	0.894	0.949	0.968	0.948	0.904	0.959	0.990

Table 4. The comparative ablation studies of HCIA evaluated in single domain on Kvasir-SEG and CVC-ClinicDB. *Base* denotes the adopted baseline. *H* denotes HAM while *I* denotes IAM.

Effectiveness of each element. The edge that HAM brings to the table is evident when examining the leap in performance observed from the baseline to the baseline augmented with HAM ('$Base + H$' in Table 3). This manifests as gains in *mDice* by 1.8%, and 1.3% on Kvasir-SEG and CVC-ClinicDB, respectively. The HAM module evidently excels at mitigating background distractions and accentuating foreground entities, enabling more precise segmentation demarcations. Conversely, IAM's contribution is highlighted by the notable upticks in segmentation accuracy upon its integration into the baseline model ('$Base + I$' in Table 3), validating its role in fostering coherent context dependencies and centralized global supervision. This integration yields improvements of 2.3%, and 2.1% in *mDice* metric across the datasets. It can be observed that, owing to the linear complexity design of the IAM, the increase in both Params and FLOPs introduced by IAM is minimal. This characteristic endows the HCIA with a significant advantage over other methods that utilize attention mechanisms. When both HAM and IAM are harmonized within a single framework, HCIA unleashes its full potential, culminating in substantial improvements of 3.7% and 3.3% in *mDice* when benchmarked against the baseline model over all three datasets ('$Base + H + I$' in Table 3). This composite effect underscores HCIA's proficiency in managing features across multiple scales, seamlessly establishing discriminative context dependencies and integrated global range interdependencies.

Effectiveness of Attention Mechanism Design in IAM. The distinctive architecture of the IAM in the proposed HCIA is scrutinized through comparative ablation experiments to validate the impact of its design elements, particularly the use of external attention and the global sharing of memories (M_k and M_v). Traditional attention mechanisms like spatial attention (SA) and channel attention (CA) are generally employed to augment the capture of localized details within images. Yet, the segmentation of complex structures, such as polyps, demands more than just local awareness—it necessitates an ability to also grasp the broader, long-range semantic context that SA and CA alone may inadequately provide due to the limits of their design. When examining the performance across the variants detailed in Table 4, specifically those incorporating attention—-I^{∇} with SA, $I^{\triangle}$ with CA, and $I^{\square}$ with both—it becomes apparent that the IAM configuration employing external attention ($Base + H + I$) surpasses each of these. The metrics demonstrate that $Base + H + I$ outperforms the best of these individual configurations, $Base + A + I^{\square}$, by a margin of 0.7% and 0.7% in $mDice$ across all datasets. This reinforces the notion that external attention is critical in capturing the complexities of polyp segmentation. Furthermore, HCIA's design introduces IAMs across four hierarchical branches, wherein each module refines the features at their respective scale to fortify the local context connections. The collective insights and oversight among the IAMs are facilitated by global memories, which strengthen the coordination and global supervision of the entire system. In the $I^{\diamond}$ variant, where M_k and M_v are not shared globally but kept exclusive to each IAM, a comparative decline in performance becomes evident when measured against the IAM using globally pooled memories ($Base + A + I$). This discrepancy highlights the benefits of a shared global memory framework, confirming that the global interworking of information across IAMs significantly contributes to HCIA's state-of-the-art performance in polyp segmentation tasks.

Qualitative analysis

Visual representations are indeed compelling when it comes to demonstrating the performance and strengths of a segmentation model. By bringing forth a selection of representative methods, including the CNN-based U-Net, PraNet and UM-Net, as well as the Transformer-based Transfuse, Polyp-LVT and CAFE-Net, for a side-by-side visual comparison with the proposed HCIA, we gain clear insights into the practical capabilities of the methodologies. From the snapshots provided in Fig. 6, the effectiveness of the HCIA, is evident across a range of challenging polyp segmentation cases on Kvasir-SEG. HCIA adapts adeptly to polyps with diverse sizes, shapes, and textures–even when boundary delineation is made difficult by varying lighting conditions that obscure the difference between the polyp tissues and surrounding normal tissues. HCIA's proficiency in these scenarios underscores the necessity for models to possess robust local information extraction tools coupled with a comprehensive global perspective—qualities endowed by the integration of HAM and IAM within HCIA. In scenarios exhibited by Fig. 7, presented with CVC-ClinicDB dataset images, where the foreground-background contrast is accentuated by improved lighting conditions, it is apparent that while most of the compared methods achieved relatively high accuracy in segmentation tasks, HCIA consistently excels at drawing detailed and precise segmentation contours. This is particularly noteworthy in cases where polyps display intricate edges as seen in rows (3) of Fig. 7. HCIA's capability to delineate complex polyp edges with a high degree of accuracy not only speaks to its performance but also to its versatility in adapting to different imaging conditions and the variable nature of polyp structures. Through the visual comparison of segmentation results under differing conditions, the HCIA's efficiency in processing polyps of various scales, shapes, and complex contours is affirmed.

Efficiency evaluation

To demonstrate the superior efficiency of the proposed HCIA, we conducted a comprehensive comparative analysis of model efficiency and complexity on the Kvasir-SEG dataset, as illustrated in Table 5. We utilized floating point operations (FLOPs), network parameters (Params) and frames per second (FPS) as evaluation criteria. It is evident that HCIA exhibits a balanced count of Params and FLOPs while achieving improved segmentation performance. Polyp-PVT demonstrates optimal model efficiency, with 25.1M Params and 11.2G FLOPs while our HCIA only incurs an increase of 0.7M Params and 2.5G FLOPs, yielding enhancements of 2.5% in *mDice* and 2.7% in *mIoU*. The comparative results validate that HCIA achieves an effective balance between efficiency and performance. Additionally, HCIA's inference speed reaches 51 FPS, meeting the requirements for real-time predictions.

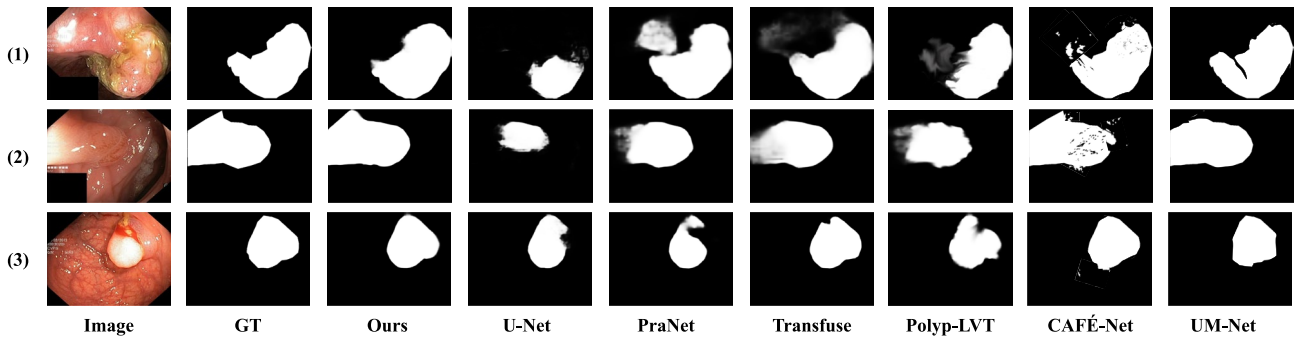

Fig. 6. Qualitative analysis on Kvasir-SEG. GT denotes the ground truth.

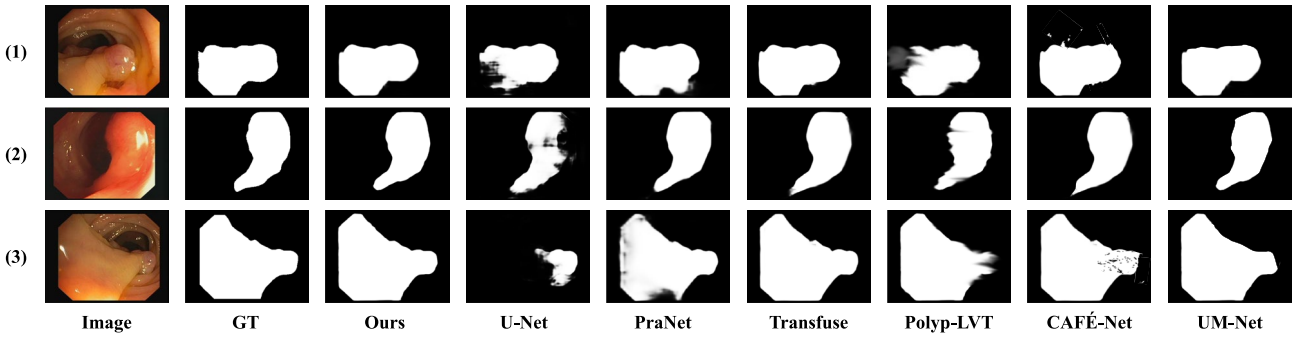

Fig. 7. Qualitative analysis on CVC-ClinicDB. GT denotes the ground truth.

Methods	Year	Type	*mDice*	*mIoU*	Params	FLOPs	FPS
U-Net	2015	CNN	0.815	0.742	31.0	103.5	54
U-Net++	2018	CNN	0.817	0.740	47.2	377.5	69
PraNet	2020	CNN	0.859	0.844	30.5	13.2	53
Polyp-PVT	2021	Transformer	0.917	0.862	**25.1**	**11.2**	53
Transfuse	2021	Transformer	0.920	0.871	26.2	21.8	-
SSFormer	2022	Transformer	0.935	0.890	29.3	19.1	-
APCNet	2023	CNN	0.913	0.859	33.1	16.3	-
SRaNet	2023	CNN	0.921	0.870	24.9	-	-
MGCBFormer	2023	Transformer	0.933	0.887	103.4	91.1	-
PVT-CASCADE	2023	Transformer	0.924	0.875	35.3	15.4	-
CTNet	2024	Transformer	0.917	0.863	44.2	15.2	-
UM-Net	2025	CNN	0.930	0.882	22.8	15.6	50
CTHP	2024	Transformer	0.939	0.891	47.1	54.2	-
Polyp-LVT	2024	Transformer	0.909	0.851	25.1	13.2	-
CAFE-Net	2024	Transformer	0.933	0.889	35.5	16.1	-
Polyp-Mamba	2025	CNN	0.919	0.867	49.5	27.9	-
EFA-Net	2025	CNN	0.914	0.861	27.4	33.2	-
WBANet	2025	Transformer	0.933	0.889	38.5	11.8	-
HCIA(ours)	-	Transformer	**0.942**	**0.894**	25.8	13.7	51

Table 5. Efficiency evaluation ofHCIA compared to other SOTA models. Optimal results are highlighted in bold.

Limitations

Despite the promising performance of HCIA, there are several limitations worth noting. First, although the IAM is designed with linear complexity to improve efficiency, the multi-scale feature aggregation in HAM and the hierarchical architecture still involve a considerable number of parameters. This may pose challenges for deploying the full model on strictly resource-constrained edge devices (e.g., embedded endoscope processors) for real-time inference without further optimization like model quantization or pruning. Second, the current framework operates on a frame-by-frame basis (2D) and does not yet leverage the temporal consistency available in colonoscopy video sequences, which could potentially further enhance detection stability.

Conclusion

In this paper, we proposed a novel Hierarchical Contextual Information Aggregation Network (HCIA) for accurate polyp segmentation. To address the challenge of balancing global context and local details, we introduced two key components: Interconnected Attention Module (IAM) and Hierarchical Aggregation Module (HAM). Specifically, IAM captures long-range dependencies across all layers via a globally shared memory mechanism with linear complexity, while HAM effectively integrates features from adjacent levels to enhance multi-scale representation. Comprehensive experiments on multiple benchmarks demonstrate that HCIA achieves state-of-the-art performance, exhibiting superior accuracy and generalization capabilities compared to existing methods.

Data availability

The Kvasir-SEG dataset analysed during the current study is available in the Kvasir-SEG repository, https://datasets.simula.no/kvasir-seg/and the CVC-ClinicDB dataset analysed during the current study is available in the CVC-ClinicDB repository, https://polyp.grand-challenge.org/CVCClinicDB/.

Received: 11 July 2025; Accepted: 7 January 2026
Published online: 14 January 2026

References

1. Siegel, R. L., Miller, K. D., Wagle, N. S. & Jemal, A. Cancer statistics, 2023. *Ca Cancer J Clin*. **73**, 17–48 (2023).
2. Kim, N. H. et al. Miss rate of colorectal neoplastic polyps and risk factors for missed polyps in consecutive colonoscopies. *Intestinal research*. **15**, 411–418 (2017).
3. Lee, J. et al. Risk factors of missed colorectal lesions after colonoscopy. *Medicine*. **96** (2017).
4. Ronneberger, O., Fischer, P. & Brox, T. 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*, 234–241 (Springer, 2015).
5. Zhou, Z., Rahman Siddiquee, M. M., 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: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, 3–11 (Springer, 2018).
6. Jha, D. et al. Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE international symposium on multimedia (ISM)*, 225–2255 (IEEE, 2019).
7. Vaswani, A. et al. Attention is all you need. *Advances in neural information processing systems*. **30** (2017).
8. Jha, D., Tomar, N. K., Sharma, V. & Bagci, U. Transnetr: transformer-based residual network for polyp segmentation with multi-center out-of-distribution testing. In *Medical Imaging with Deep Learning*, 1372–1384 (PMLR, 2024).
9. Brandao, P. et al. Fully convolutional neural networks for polyp segmentation in colonoscopy. In *Medical Imaging 2017: Computer-Aided Diagnosis*, vol. 10134, 101–107 (Spie, 2017).
10. Fan, D.-P. et al. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, 263–273 (Springer, 2020).
11. Tomar, N. K. et al. Ddanet: Dual decoder attention network for automatic polyp segmentation. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VIII*, 307–314 (Springer, 2021).
12. Zhao, Y., Li, J. & Hua, Z. Tact: Text attention based cnn-transformer network for polyp segmentation. *International Journal of Imaging Systems and Technology*. **34**, e22997 (2024).
13. Huang, Z. et al. Mgf-net: Multi-channel group fusion enhancing boundary attention for polyp segmentation. *Medical Physics*. **51**, 407–418 (2024).
14. Liu, Y., Shen, X., Lyu, Y. & Wang, X. Mca-net: multi-cascade attention network for polyp segmentation. *Multimedia Tools and Applications*. **83**, 33713–33730 (2024).
15. Du, X. et al. Um-net: Rethinking icgnet for polyp segmentation with uncertainty modeling. *Medical Image Analysis*. **99**, 103347 (2025).
16. Zhu, X., Wang, W., Zhang, C. & Wang, H. Polyp-mamba: A hybrid multi-frequency perception gated selection network for polyp segmentation. *Information Fusion*. **115**, 102759 (2025).
17. Wang, J. et al. Stepwise feature fusion: Local guides global. arxiv 2022. arXiv:2203.03635 .
18. Duc, N. T., Oanh, N. T., Thuy, N. T., Triet, T. M. & Dinh, V. S. Colonformer: An efficient transformer based method for colon polyp segmentation. *IEEE Access***10**, 80575–80586 (2022).
19. Zhang, Y., Liu, H. & Hu, Q. Transfuse: Fusing transformers and cnns 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 I 24*, 14–24 (Springer, 2021).
20. Chen, J. et al. Transunet: Transformers make strong encoders for medical image segmentation. arXiv:2102.04306 (2021).
21. Dong, B. et al. Polyp-pvt: Polyp segmentation with pyramid vision transformers. arxiv 2021. arXiv:2108.06932 .
22. Xue, H., Yonggang, L., Min, L. & Lin, L. A lighter hybrid feature fusion framework for polyp segmentation. *Scientific Reports*. **14**, 23179 (2024).
23. Liu, G. et al. Cafe-net: Cross-attention and feature exploration network for polyp segmentation. *Expert Systems with Applications*. **238**, 121754 (2024).
24. Xiao, B., Hu, J., Li, W., Pun, C.-M. & Bi, X. Ctnet: Contrastive transformer network for polyp segmentation. *IEEE Transactions on Cybernetics*. **54**, 5040–5053 (2024).
25. Wang, Y., Tian, Q., Chu, J. & Lu, W. A wavelet-enhanced boundary aware network with dynamic fusion for polyp segmentation. *Neurocomputing*. 130259 (2025).
26. Zhang, W. et al. Hsnet: A hybrid semantic network for polyp segmentation. *Computers in biology and medicine*. **150**, 106173 (2022).
27. Liu, Y., Yang, Y., Jiang, Y. & Xie, Z. Multi-view orientational attention network combining point-based affinity for polyp segmentation. *Expert Systems with Applications*. **249**, 123663 (2024).
28. Wang, W. et al. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*. **8**, 415–424 (2022).
29. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*. **25** (2012).
30. Oktay, O. et al. Attention u-net: Learning where to look for the pancreas. arXiv:1804.03999 (2018).
31. Guo, M., Liu, Z., Mu, T. & Hu, S. Beyond self-attention: External attention using two linear layers for visual tasks. arxiv 2021. arXiv:2105.02358 (2021).
32. 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*, 451–462 (Springer, 2020).
33. Bernal, J. et al. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*. **43**, 99–111 (2015).
34. Tajbakhsh, N., Gurudu, S. R. & Liang, J. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*. **35**, 630–644 (2015).
35. Shi, J.-H., Zhang, Q., Tang, Y.-H. & Zhang, Z.-Q. Polyp-mixer: An efficient context-aware mlp-based paradigm for polyp segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*. **33**, 30–42 (2022).
36. Milletari, F., Navab, N. & Ahmadi, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, 565–571 (Ieee, 2016).
37. Cheng, M.-M. & Fan, D.-P. Structure-measure: A new way to evaluate foreground maps. *International Journal of Computer Vision*. **129**, 2622–2638 (2021).
38. Fan, D.-P. et al. Enhanced-alignment measure for binary foreground map evaluation. arXiv:1805.10421 (2018).
39. Ma, J. et al. Segment anything in medical images. *Nature Communications***15**, 654 (2024).

40. Yin, Z., Liang, K., Ma, Z. & Guo, J. Duplex contextual relation network for polyp segmentation. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 1–5 (IEEE, 2022).
41. Nam, J.-H., Syazwany, N. S., Kim, S. J. & Lee, S.-C. Modality-agnostic domain generalizable medical image segmentation by multi-frequency in multi-scale attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11480–11491 (2024).
42. Muhammad, U. et al. Mmfil-net: Multi-level and multi-source feature interactive lightweight network for polyp segmentation. *Displays*. **81**, 102600 (2024).
43. Lin, L., Lv, G., Wang, B., Xu, C. & Liu, J. Polyp-lvt: Polyp segmentation with lightweight vision transformers. *Knowledge-Based Systems*. **300**, 112181 (2024).
44. Yue, G. et al. Attention-guided pyramid context network for polyp segmentation in colonoscopy images. *IEEE Transactions on Instrumentation and Measurement*. **72**, 1–13 (2023).
45. Zhou, T. et al. Edge-aware feature aggregation network for polyp segmentation. *Machine Intelligence Research*. **22**, 101–116 (2025).
46. Dong, B. et al. Polyp-pvt: Polyp segmentation with pyramid vision transformers. arXiv:2108.06932 (2021).
47. Lee, G.-E., Cho, J. & Choi, S.-I. Shallow and reverse attention network for colon polyp segmentation. *Scientific Reports*. **13**, 15243 (2023).
48. Yue, G. et al. Boundary uncertainty aware network for automated polyp segmentation. *Neural Networks*. **170**, 390–404 (2024).
49. Wang, J. et al. Stepwise feature fusion: Local guides global. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 110–120 (Springer, 2022).
50. Wang, H. et al. Dynamic spectrum-driven hierarchical learning network for polyp segmentation. *Medical Image Analysis*. **101**, 103449 (2025).
51. Xia, Y., Yun, H., Liu, Y., Luan, J. & Li, M. Mgcbformer: The multiscale grid-prior and class-inter boundary-aware transformer for polyp segmentation. *Computers in Biology and Medicine*. **167**, 107600 (2023).
52. Sanderson, E. & Matuszewski, B. J. Fcn-transformer feature fusion for polyp segmentation. In *Annual Conference on Medical Image Understanding and Analysis*, 892–907 (Springer, 2022).
53. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. arXiv:1412.6980 (2014).
54. Chen, L.-C., Papandreou, G., Schroff, F. & Adam, H. Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587 (2017).
55. Srivastava, A. et al. Msrf-net: a multi-scale residual fusion network for biomedical image segmentation. *IEEE Journal of Biomedical and Health Informatics*. **26**, 2252–2263 (2021).

Funding

We state that there was no Funding for this manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to X.H.