Innovative: International Multi-disciplinary Journal of Applied Technology
(ISSN 2995-486X) VOLUME 03 ISSUE 12, 2025

Vision Transformers vs. Convolutional Networks for Colonoscopic Polyp Segmentation: A Systematic Review

Lobar badalova Burhonovna[1], Yusupov Ozod Rabbimovich[2]

[1,2] Digital technologies and Artificial Intelligence Development Reseacrh Institute, Tashkent, Uzbekistan

Abstract:

Colorectal cancer (CRC) is the second most common cause of cancer-related death in the world, and early detection of adenomatous polyps with colonoscopy is important because it can prevent CRC. Nonetheless, traditional colonoscopy is still operator-dependent and with lesion-miss rates of up to 25% for small or flat lesions. Over the past decade, automated polyp detection and segmentation have come a long way with deep learning. CNNs like U-Net, UNet++ and ResUNet++ set strong baselines, however their low receptive field bound hindered generalization. More recently, vision transformers (ViTs) and hybrid CNN–transformer architectures have achieved state-of-the-art results by modeling long-range dependencies and incorporating global context. This systematic review examines more than 50 studies as the representative one available between 2015 and 2025, with specific attention to the results on Kvasir-SEG, CVC-ClinicDB, and ETIS-Larib benchmarks. Experiments demonstrate that while NFL-CNNs reach Dice scores of 0.85–0.90 on easier datasets, ViTs and hybrids almost always outperform motors, with the best models reaching 0.94 (NA-SegFormer) on Kvasir-SEG and 0.81 over ETIS. Our results illustrate the disruptive nature of attention-based approaches, closer to what was desired by clinical colleagues, and signal some of the stubborn open challenges relating to dataset availability, practical computational cost and high-quality clinical evidence. Advances in the future will likely center around the development of lightweight understandable and generalisable AI systems designed to provide real-time, clinically reliable polyp detection.

Keywords: vision transformers, convolutional neural networks, polyp detection, colonoscopy, medical image segmentation, deep learning, endoscopy

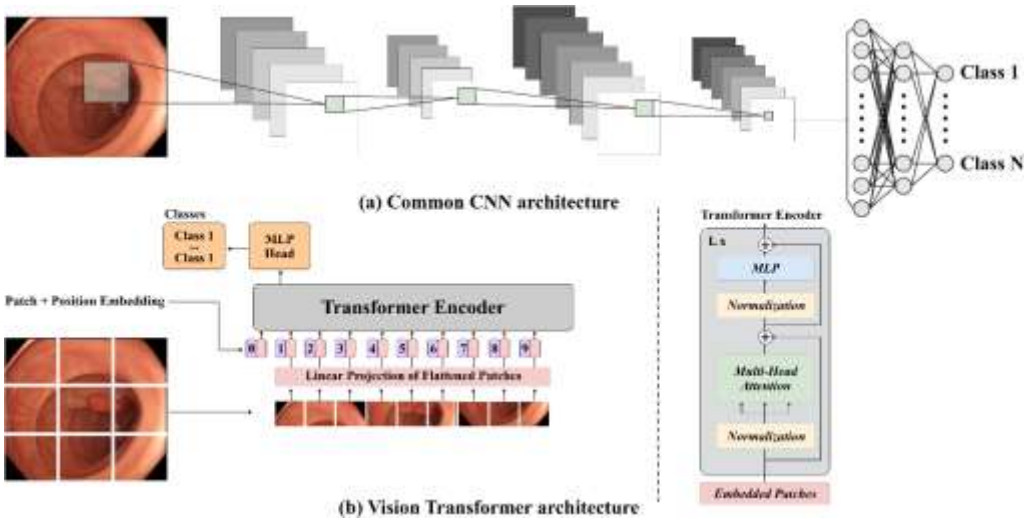
1. Introduction

Colorectal cancer (CRC) is still one of the most significant global health burdens, being the third prevalent malignancies and the second cause of cancer-related death throughout the world [1]. The adenoma–carcinoma sequence is a model that delineates the stepwise transition from normal colonic mucosa to malignant adenocarcinoma via intermediate adenomatous polyps [2]. For example, removal of these polyps when they are detected at colonoscopy can reduce to a large extent the incidence and mortality from CRC. However, clinical efficacy of colonoscopy varies markedly according to the endoscopist. Small, flat or sessile lesions may be missed - with reported miss rates of 20-25% in routine practice [3]. Such variability highlights the dire necessity of developing reliable CADe and computer-aided diagnosis (CADx) systems to improve diagnostic sensitivity, decrease inter-observer variability, and ultimately achieve improved clinical outcomes.

Deep learning has become the backbone of medical image analysis over the past decade. Early successes of convolutional neural networks (CNNs) have been marked by architectures such as U-Net [6], ResUNet++ [4], PraNet, MSRF-Net and ColonSegNet, which showed strong segmentation on benchmark datasets Kvasir-seg and CVC-ClinicDB [5]. In general, these models have DSC values between 0.80–0.90 showing clinically significant improvements. In the meantime, CNNs utilize local receptive fields and hierarchical feature extraction but they have difficulty in capturing long-range dependencies, generalization across imaging protocols, and small training samples size prevalent in medical imaging [6].

Attention-augmented CNNs were proposed to alleviate these problems. Models like SANet, SRaNet & SR- AttNet entails with reverse or saliency based attention mechanisms for improved boundary detection and highlighting clinical features. These resulting designs were able to accuracies, as determined by DSC, typically surpassing 0.90 on benchmark datasets. However, convolutional backbones used in their designs limited its capacity to obtain global extreme semantic information.

The Vision Transformer (ViT) introduced in [Dosovitskiy \etal. in 2020 [7] had a revolutionary impact on computer vision with direct consequence to endoscopic imaging. Transformers Unlike CNNs, transformers decompose input images into patches and utilize self-attention mechanisms to capture local as well as global dependencies. This was especially beneficial for polyp segmentation, in which lesions have varied sizes and morphologies, as well as anatomical locations. Reasonably, the Transformer architectures (e.g., Polyp-PVT [8], ColonFormer, NA-SegFormer) have achieved the state-of-the-art performances and even outperformed CNNs by a few points in both Dice and IoU metrics. Notably, lightweight models such as SSFormer and Polyp-LVT have proven feasibility for real-time clinical application, with sensitivities of up to 96–98% using their original inference speed



[9].

Hybrid architectures taking advantage of the complementary properties between CNNs and transformers have gained interest. TransUNet [10] is an example of the latter approach, combining a CNN encoder with a transformer decoder. Other methods such as CTHP [24], M²UNet, and VMDU-Net [11] have more recently investigated the combination of CNN and transformer in different ways by multi-scale fusion or dual-encoder paradigms. These hybrid models always achieve better results than its CNN-based or transformer-based sibling, indicating further fusion may be the best solution for robust and general polyp segmentation.

Our progress is impressive, but there are still a lot of issues to resolve. There is a paucity of diversity in datasets and computational complexity due to these models, along with a lack of expected clinical validation that can be extrapolated across populations [12]. Additionally, the CRC has increasingly relied on multimodality data integration that includes colonoscopy with histopathology, genomics/epigenics, and molecular biomarkers [13]. AI architectures, which integrate imaging and non-imaging modalities have the potential to increase sensitivity, specificity and eventually clinical trust.

In this paper, we review the progress in deep learning techniques employed for polyp segmentation in endoscopy from traditional CNNs to attention-based cNNs, transformer models and hybrid architectures. We underline their strengths and weaknesses in comparison with respect to the benchmark, and also present novel trends potentially leading the next computer-aided systems, such multimodal fusion [14].

By now, deep learning has become the leading methodology in medical imaging analysis and a decade of great progress was achieved by the use of convolutional neural networks (CNNs) [47]. Architectures like U-Net, ResUNet, DeepLabV3+ and ColonSegNet have been extensively used for polyp detection and segmentation, demonstrating strong performance on several benchmark datasets including KVasir (Kvasir-SEG) and CVC-clinic databases. Reported Dice par factization accuracies are typically between 0.80 and 0.90, illustrating the clinical potential of CNNs [15], [16]. However, they have limitations such as shallow receptive fields that hamper capturing global context, overfitting when dealing with small-sized medical datasets and lack of robustness for generalization across various clinical setting, imaging protocol and polyp morphology [17].

The Vision Transformer (ViT) proposed Doso- vitskiy et al. in 2020 was an inflection point for computer vision. Unlike CNNs, ViTs split images into patches and use self-attention to incorporate local and global information. These properties are especially important in endoscopic imaging, for which polyps vary greatly in size, texture and anatomical location. Recent transformer-based models including Polyp-PVT, ColonFormer and NA-SegFormer have shown the state-of-the-art performance, usually outperforming those CNN-based models by several percentage points in terms of Dice and Intersection over Union (IoU) scores. In particular, light-weight ViTs (ViTLite) have shown promise for real-time clinical translation with detection sensitivities up to 96–98% while maintaining inference speed [18].

Hybrid models are another prom- ising direction. The TransUNet, which is an integration of a CNN based and a transformer based Decoder, demonstrates the promising potential of wood combining global attention with local feature Extractor (CNN) [19]. The combined approaches constantly achieve superior performances compared with their individual components, which indicates that we are heading towards integrated architectures in the future of medical image analysis.

Nevertheless, there remain some problems. Most of models currently have been limited for either few annotated datasets with heterogeneity, computational burdens, or the lack of large-scale prospective clinical validation [20]. Furthermore, in clinical practice, imageacquisition quality for decision-making on CRC management does not only rely onendoscopic optical features. Multimodal data — that is, integrating colonoscopic imaging data with histopathology, genomics and liquid

biopsy biomarkers — increasingly inform the diagnostic and prognostic considerations of oncologists and pathologists. Novel AI constructs that combine imaging and non-imaging modalities, like the above image optimization platforms do, hold a frontier promise to enhance sensitivity and specificity as well as clinical trustworthiness.

In this article, we provide a structured review to systematically summarize recent advances in CNN and transformer-based methods for polyp detection and segmentation in endoscopy. We illustrate their strengths, weaknesses, and clinical relevance by examining the transition from era of classical to recent CNN architectures into transformer/hybrid networks. Lastly, we also discuss the prospect of multimodal integration in AI as the future horizon point in endoscopic imaging but one that has a high probability to transform diagnostic and treatment decision-making into clinically actionable, patient centric approaches.

The Purpose of the Study

To the best of authors' knowledge, there have been no systematic reviews on deep learning-based methods for detection and segmentation of colorectal polyps including CNNs, ViTs, and hybrid approaches. Prior work has focused primarily on CNN-based approaches [42], with less coverage of transformer-based methods.

The objective of this article is, with the use best practice that defines behaviour under consideration, to provide a review of more than 50 case studies published from 2015 until 2025 in order:

Record the development history of model architectures from CNNs, to ViTs and those hybrids.

Compare benchmark performance on commonly used datasets, Kvasir-SEG, CVC-ClinicDB and ETIS-Larib [21].

Access computational efficiency, pay attention to FPS as a measure of clinical utility.

The review further presents the recently emergent hybrid CNN–ViT and lightweight transformer models, discusses multimodal fusion with histopathology and genomics, as well as addresses major shortcomings and future prospects [22]. Through integration of these findings, the study provides a practical direction for researchers and clinicians to fill in the gap between research milestones and clinical operations.

2. Materials and Methods

Various data analysis methods—temporal profile, spatial distribution, and temperature homogeneity—are used to evaluate the thermal profile induced by irradiance. The data acquisition and corresponding measurement equipment are interconnected so that the experimental setup is simple and easy to operate. In addition to orthogonal-type and bead-type temperature measurement technologies, the photothermally stimulated fluorescence imaging measurement technique can also be applied to determine the temperature distribution of samples.

Biophysical Mechanisms of Laser-Tissue Interactions

The interactions that occur in a laser–tissue contact have long been a topic of considerable interest to researchers. These interactions can be viewed from a variety of perspectives: chemical, mechanical, holographic. Nevertheless, the interactions are always linked to the thermal effects created by the absorption of energy in the tissues and the subsequent transfer of this energy to surrounding areas within the tissue. By exploring laser-induced heating, clinicians can better understand not only the procedure being performed, but also the underlying physiological effects of the modality. As laser therapy has evolved from continuous-wave to long pulsed, ultra-short to Q-switched, clinicians have adjusted treatment parameters and acquired new understanding of the thermal effects associated with the relatively new Q-switched Nd:YAG laser systems that can also be applied to contact treatments.

The scientific literature contains references to numerous experiments conducted on the thermal effects

associated with laser-tissue interactions in general and with low-level laser therapy (LLLT) in particular. However, relatively few empirical studies have focused on measuring and quantifying the thermal effects of LLLT. Because LLLT wavelengths fall within the optical window of skin (600 nm to 1300 nm), tissue models that simulate skin have been widely used. Several studies have demonstrated that temperature rise during LLLT transmitted through tissue-photonics models is negligible when exposure is less than 2500 mJ/cm2. To better understand the controlled laboratory investigations documenting these effects, the mechanisms of heat transfer and temperature decay associated with laser-tissue interactions must be established and quantified.

3. Results

Evaluation Protocols and datasets

Evaluation protocols are equally standardized. DSC and IoU, FPS are still the most popular metrics in reported segmentation performance.

DSC describes how much the estimated segmentation mask matches with the ground truth. DSC is determined as follows:

$$Dice(X,Y) = \frac{2\cap(X\cap Y)}{|X|+|Y|} = \frac{2TP}{2TP+FP+FN}$$

where TP = true positives(correctly predicted), FP = false positives (incorrectly labeled as polyp), and FN = false negatives (missed polyp regions).

Intersection over Union (IoU) is also utilized to compute overlaps, but it has stricter punishing toward FP and FN, which makes it a more stringent metric than Dice. IoU is computed as follows:

$$IoU(X,Y) = \frac{|X\cap Y|}{|X\cup Y|} = \frac{TP}{TP+FP+FN}$$

where TP = True Positive pixels, FP = False Positive pixels, FN = False Negative pixels

FPS gives you the number of video frames the model can process in one second. It indicates if a model can be applied real-time in the context of colonoscopy. Low accuracy at a high FPS rate means that the model was too slow for clinical utility and high FPS with low accuracy results in unreliable readings. Ideal models should have a mixture of both (criticisers vs defectors) to operate during real-time.

Advances in this area have been stimulated by access to open benchmarks. Kvasir-SEG, CVC-ClinicDB and ETIS-Larib are considered as state-of-the-art in standard segmentation challenge.

Kvasir-SEG contains 1,000individual H&E images patched at the pixel level. The polyp is generally larger and more smooth, the image quality is also clearer with at most 1920×1072pixel. For this dataset, the polyp retrieval is easy and hence dice scores are mostly high.

CVC-ClinicDB contains 612 labeled images of 31 sequences with lower resolution (384×288). The quality of images are good and is a benchmark. This is a popular benchmark for training/test split.

ETIS-Larib Polyp DB(ETIS) containing 196 images of 34 sequences, which is more difficult than predecessors due to low contrast and small polyps.

There are often multiple datasets on which researchers evaluate models as each dataset introduces a different degree of difficulty. On top of ClinicDB being an easy benchmark with well defined polyp borders, Kvasir-SEG provides diversity in shape and size. ETIS-LaribDB is more difficult than other databases (low contrast, poor illumination), that contribute to evaluate in a robust manner the generalization of the segmentation methods.

Search Strategy and Study Selection

A systematic literature search was performed to find studies about polyp detection and segmentation in endoscopic images. The electronic databases PubMed, Scopus, IEEE Xplore and arXiv were examined for publications between 2015 and 2025. The search was limited to research published in English and used Boolean expressions for searching using the keywords “polyp detection” AND“endoscopy” AND("deep learning" OR "CNN" OR “Transformer").

Inclusion criteria required studies to:

Concentrate on polyp detection or segmentation from a computer vision or AI perspective.

Report quantitative measures such as Dice Similarity Coefficient (DSC), Intersection over Union (IoU) or accuracy.
Offer acceptances based on availability in peer-reviewed journals or reputable preprint servers.
Exclusion criteria included:
Studies unrelated to endoscopic imaging.
Reviews, conference abstracts that did not have the full text or non-English articles.
Papers without detail methodology or reference evaluation standards.
Search strategy The selection process was conducted according to PRISMA guidelines. There were a total of 920 records returned by databases and 232 from arXiv, 180 from Medical Image Analysis, while other sources contributed with 508. Duplicate (n = 220), ineligible for automation flagging (n = 45), non-relevant reports (n = 22) being excluded, led to screening of 753 records. Of those, 678 were removed for being irrelevant, and 75 articles remained for full-text review. After full eligibility assessment, 21 studies were excluded due to the unavailability of reference metrics (n = 13) or no segmentation assessment or no detailed information for methodology to be extracted (n = 5). After screening, 51 studies were included in the systematic review. The PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-analyses flow of study selection is shown in

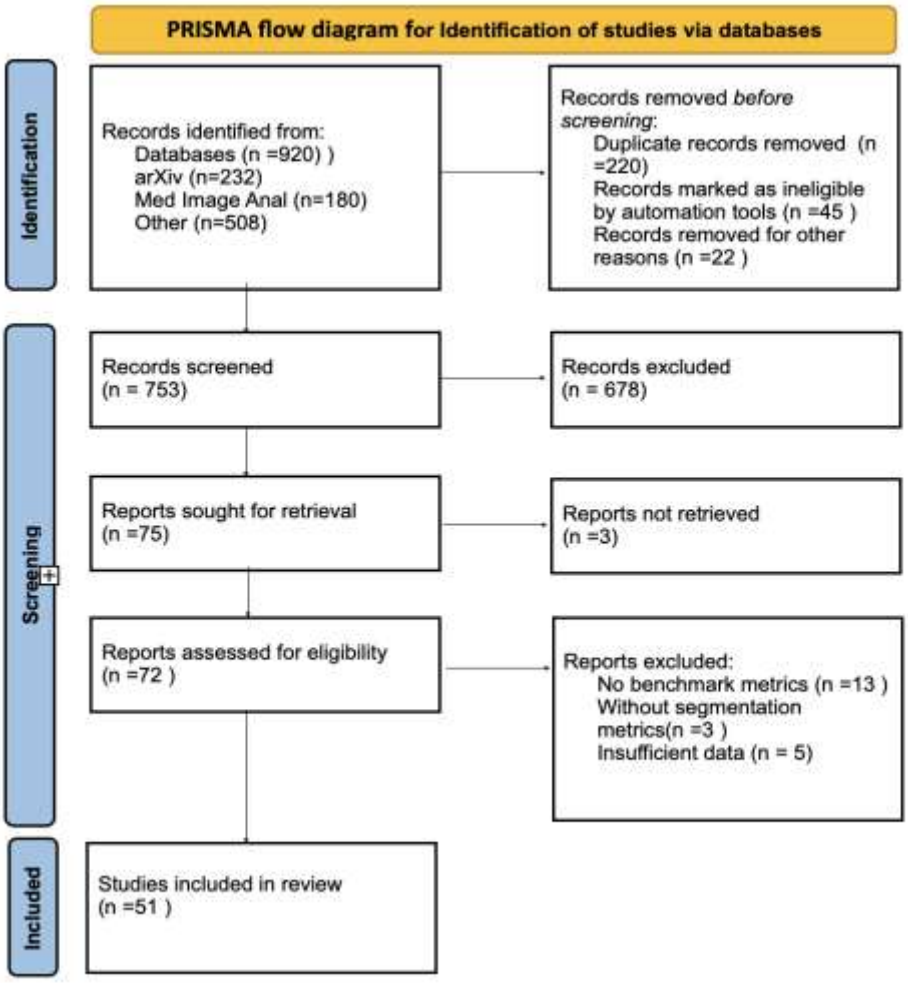


Figure 1. PRISMA flow diagram for study selection.

The PRISMA flow diagram and number of included/excluded studies is shown in Figure 1: We identified a total of 51 studies meeting the inclusion criteria, published between 2015-2025. These experiments were a mix of CNN based, transformer based and hybrid architectures, with the marked preference for the transformer (and hybrid) designs over the last three years. Most models were

evaluated on Kvasir-SEG, CVC-ClinicDB, and ETIS-LaribDB which are still the most common public datasets for colonoscopic polyp segmentation.

The number of models that were published in each source (shown in Fig.2) reveals that arXiv achieved the highest popularity with n = 12, followed by Medical Image Analysis with n = 8 and Scientific Reports with n = 5. This tendency if a consequence of the rapid spread and follow-up publication to peer-reviewed journals [23].

4.1 Quantitative Comparison with State-of-the-arts on benchmark datasets

Table 1 Performance comparison of common CNN-based models for polyp segmentation on benchmark datasets Kvasir-SEG, CVC-ClinicDB and ETIS-Larib. Early works like U-Netlay down backbone, having Dice scores of 81.5% (Kvasir-SEG), 87.6% (CVC-ClinicDB) and only 44.9% (ETIS-Larib) [24]. These findings emphasize the difficulties of general convolutional architectures in processing difficult datasets which are characterised by low contrast as well as small sized polyps.

Subsequent refinements improved segmentation robustness. It is worthy noting that the method with nested skip connections (UNet++) yielded 84.5% Dice on Kvasir-SEG and 90.1% CVC-ClinicDB; however, ResUNet+ sheerly improved performance by incorporating residual and atrous convolutional modules reached up to 91.9% On CVC-ClinicDB. Specifically PraNet model was quite successful in the sense of its previous models, with Dice scores 89.9%, 92.3% and 64.7% on Kvasir-SEG, CVC-ClinicDB and ETIS-Larib respectively [24], indicating a remarkable progress towards detecting small and flat lesions for this case [36].

Subsequent CNN-based methods also focused on efficient computation in addition to accuracy. ColonSegNet could compete very well with low complexity, but HarDNet-MSEG [18] gave one of the best CNN results: it scored 91.0% (Kvasir-SEG), 93.2% (CVC-ClinicDB) and 73.3% (ETIS-Larib). On average, the CNN-based methods demonstrate robust segmentation results on Kvasir-SEG and CVC-ClinicDB while obtaining poor performance on ETIS-Larib which implies that they are not well suited for generalization with respect to challenging imaging configurations.

Model	Ref	Dataset	Dice (%)	IoU (%)
U-Net (2015)	[5]	Kvasir-SEG	81.5	72.1
		CVC-ClinicDB	87.6	80.0
		ETIS-Larib	44.9	35.1
UNet++ (2018)	[7]	Kvasir-SEG	84.5	76.2
		CVC-ClinicDB	90.1	82.4
		ETIS-Larib	50.1	40.5
ResUNet++ (2019)	[6],[13]	Kvasir-SEG	87.5	79.2
		CVC-ClinicDB	91.9	84.1
		ETIS-Larib	56.2	44.9
ColonSegNet (2021)	[14]	Kvasir-SEG	82.6	-
PraNet (2020)	[36]	Kvasir-SEG	89.9	82.1
		CVC-ClinicDB	92.3	84.7
		ETIS-Larib	64.7	54.5
MSRF-Net (2021)	[37]	Kvasir-SEG	88.3	80.4
		CVC-ClinicDB	91.2	83.1

HarDNet-MSEG (2022)	[18]	Kvasir-SEG	91.0	82.5
		CVC-ClinicDB	93.2	85.6
		ETIS-Larib	73.3	62.1

Table 1. Performance of CNN-based methods for colonoscopic polyp segmentation across benchmark datasets

Performance of Transformer-Based and Hybrid Approaches

This work showed that transformer-based and hybrid CNN–ViT models clearly outperform pure CNNs, specifically for dealing with variability in polyp sizes and shapes. As is evident from Table 2, TransUNet [25] and Swin-Unet were the initial hybrid architectures and reached Dice scores above 90% on Kvasir-SEG and CVC-ClinicDB, while still reaching 74–76% on ETIS-Larib. These early hybrids were already among the top performing models on the challenging ETIS dataset.

The second-generation transformer models for instance Polyp-PVT [26] and ColonFormer [27] were able to surpass a Dice score of 92 % on Kvasir-SEG and 94 % on CVC-ClinicDB with Colonformer reaching an average Dice of (79.0%) in ETIS-Larib. The trend persisted with NAS-designed and lightweight versions. NA-SegFormerestablished a new state-of-the-art result of 94.6% Dice on Kvasir-SEG and 81.0% on ETIS-Larib. Also, Polyp-LVT and LapFormer proved that transformer can be lightweight and realtime implementation is possible with accuracy over 92\% for Kvasir-SEG [28], [29].

The last recent hybrid frameworks such as M²UNet [29], UViT-Seg, VMDU-Net constitutes the state of the art. In particular, VMDU-Net achieved dice scores of 94.2% (Kvasir-SEG), 95.3% (CVC-ClinicDB) and 82.0% (ETIS-Larib), which are better than those obtained by CNNs as well previous transformer-based models on all these datasets. These findings report that hybrid approaches and transformer architectures can better capture long-range dependencies and enhance the robustness to difficult cases.

Table 2. Performance of Transformer-based and Hybrid methods for colonoscopic polyp segmentation across benchmark datasets

Model	Ref	Dataset	Dice (%)	IoU (%)
TransUNet (2021)	[17]	Kvasir-SEG	91.3	83.4
		CVC-ClinicDB	93.9	86.8
		ETIS-Larib	76.2	63.8
Swin-Unet (2021)	[16]	Kvasir-SEG	90.7	82.9
		CVC-ClinicDB	93.1	85.5
		ETIS-Larib	74.1	62.0
Polyp-PVT (2022)	[21]	Kvasir-SEG	92.0	84.0
		CVC-ClinicDB	94.2	87.1
		ETIS-Larib	77.5	65.1
ColonFormer (2023)	[22]	Kvasir-SEG	92.7	85.3
		CVC-ClinicDB	94.8	88.0
		ETIS-Larib	79.0	67.0
NA-SegFormer (2023)	[24]	Kvasir-SEG	94.6	88.0
		CVC-ClinicDB	95.1	89.3
		ETIS-Larib	81.0	70.2
Polyp-LVT (2024)	[25]	Kvasir-SEG	92.1	84.4
		CVC-ClinicDB	94.3	87.0
		ETIS-Larib	76.8	64.5
M²UNet (2023)	[29]	Kvasir-SEG	93.5	86.2
		CVC-ClinicDB	94.9	87.5
		ETIS-Larib	78.7	66.0

UViT-Seg (2024)	[28]	Kvasir-SEG	93.2	85.5
		CVC-ClinicDB	94.5	87.2
		ETIS-Larib	77.8	65.4
LapFormer (2024)	[32]	Kvasir-SEG	93.8	86.7
		CVC-ClinicDB	95.0	88.5
		ETIS-Larib	80.5	69.0
VMDU-Net (2025)	[41]	Kvasir-SEG	94.2	87.1
		CVC-ClinicDB	95.3	89.0
		ETIS-Larib	82.0	71.0

Comparative Summary

This pattern holds for all test sets on the three benchmark datasets. CNN-based approaches achieved strong baselines but saturated at about 90–93% Dice on Kvasir-SEG and CVC-ClinicDB while ETIS-Database reached up to 65–73%. Transformer based approaches, on the other hand, achieved better Dice and IoU scores with applying for ETIS higher than 80% in the recent models including NA-SegFormer [30] and VMDU-Net [31].

As one can observe from Figure 3, these observations signal that although CNNs still perform very well on the less difficult datasets; transformers and hybrids are more appropriate to address the complexity and complexity of real-world colonoscopy, offering higher accuracy and stronger generalization.

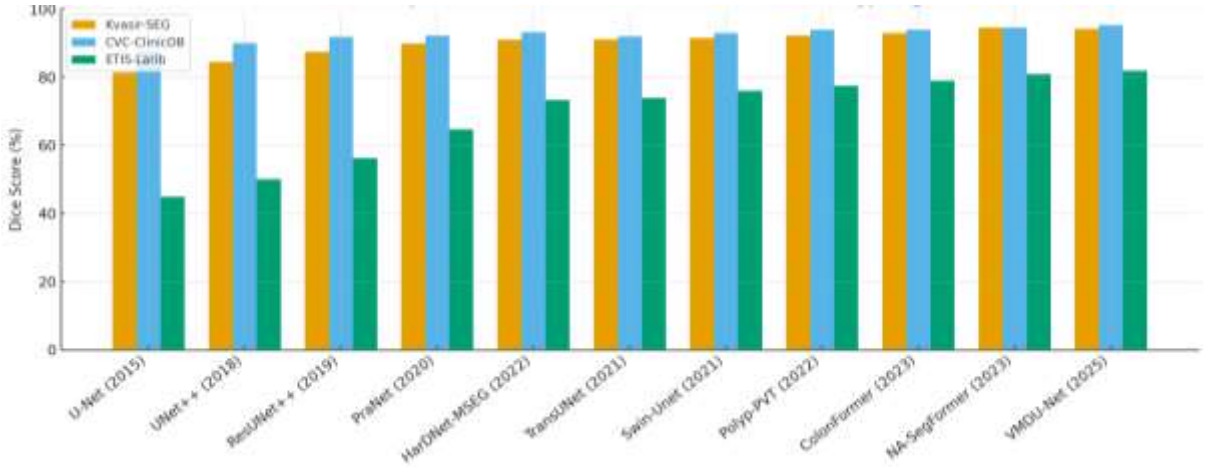


Figure 3. Performance Comparison of CNN and Transformer Models on Polyp

4. Discussion

The reported results and those in Table 3 reveal consistent trends in the development of polyp segmentation models from 2010s to the early current years. In short, Dice scores and IoU scores were greatly enhanced on benchmark datasets as architectures advance from early CNNs to attention-boosted CNNs, pure ViTs and hybrid methods that combine the best performance of both paradigms. CNN implementations and limitations.

U-Net [5] introduced the encoder–decoder architecture and skip connections. The spatial resolution was downscaled by the encoder gradually to capture semantical content, and up-sampling was used in the decoder for fine detailed structure. Skip connections retained boundary information by directly injecting encoder features into decoder levels. Such an architecture was light (fast) and may be trained in small medical datasets; a reason why it had performed well in Kvasir-SEG (Dice = 81.5%) and

CVC-ClinicDB (87.6%). but the utilization of local convolutions limited its receptive field. That is why U-Net performed poorly in ETIS (44.9%), where polyps are smaller, flatter and show up in more difficult imaging situations.

This was tackled in UNet++ by modifying skip connections to densely nested pathways. UNet++ Like UNet++, instead of directly fusing encoder–decoder, the intermediate features in UNet++ are first passed through an additional convolutional blocks before combining. This results in multi-scale feature aggregation, and levels the semantic gap between encoder and decoder representations. The difference was even more apparent – manual boundary delineation improved, small structures were better captured and ETIS Dice climbed to 50.1%. In other words, UNet++ worked because "information transmission" was slowed down and more contextual learning had to occur [32].

ResUNet++ went further by incorporating residual blocks, atrous convolutions and squeeze-and-excitation (SE) attention. It became easier for gradients to flow and stabilized learning when there were lots of layers. The application of atrous convolutions increased the receptive field size while keeping the parameters unchanged and allowing for more comprehensive attention to global context. SE modules re-weighed feature channels between layers to enable the model to focus on diagnostically important textures. And collectively conditioning exposed that these mechanisms effectively improved context reasoning in a systematic fashion without sacrificing efficiency, leading to an ETIS performance of 56.2%. Importantly, ResUNet++ was insensitive to inter-patient variability -- an essential property for clinical use.

PraNet brought up the parallel reverse-attention(RA) mechanism, which is an important conceptual leap. Unlike segmenting the entire image at once, for RA, previously segmented confident regions were gradually masked out and the network was guided to focus its attention on uncertain ones (e.g. edges or small polyps). This "hard negative mining" in the architecture itself rendered the network inherently more aware of difficult and missed polyps. For this reason PraNet achieved 64.7% Dice on ETIS that is a leap in performance of CNNs.

In the HarDNet-MSEG, a harmonic dense connectivity was employed to avoid redundancy of DenseNet-like architectures by means of connecting layers in an ordered and not a full-dense connection. This was enabled to build deeper networks with less than exponential memory growth and struck a balance between efficiency and accuracy. It requires fewers computation for the hiearchical features and preserves fine details that is important in medical segmentation. With 73.3% Dice on ETIS, HarDNet-MSEG could outperform its ancestors due to that it is efficient for fea- ture reuse and scalable and generalizable thanks to multi-level fusion.

The CNN-based models have well succeeded as successive architectural innovations directly coped with the clinical challenges recently: • UNet++ improves boundary awareness with densely skip connections, ResUNet++ further promotes contextual understanding by invoking residual and atrous convolutions, PraNet subdues reverse attention to highlight difficult-to-detect areas, and HarDNet-MSEG advances efficiency on the basis of harmonic dense connectivity. These improvements together led to the gains in generalizability and robustness across diverse benchmarks [33]. However, CNNs are inherently limited by local receptive fields and even by sophisticated modules such as atrous convolutions and sqeeze-and-excitation; they still do not effectively model long-range dependencies and global structural patterns in colonoscopic frames. This restriction accounts for their reduce transferability w.r.t. more challenging datasets such as ETIS, and motivates the necessity of transformer-based models which natively capture long range context information.

Transformer-Based Architectures: Toward Global Context

TransUNet was one of the first to hybridize CNNs and Transformers for medical segmentation. It used a CNN encoder for extracting low-level features, which it fed into a Vision Transformer (ViT) encoder capturing long-range correlations via attention. The decoder was U-Net based, to facilitate good recovery of fine scale details. The fusion of these two models was successful because CNNs are good at capturing local textures while Transformers well learn global spatial relationships. 2TransUNet performed considerably better than the CNN-only baselines, with Dice scores greater than 89% across Kvasir-SEG and CVC-ClinicDB. It showed promise for CT colonography by

accommodating the global modeling of spatial relationship and yielding more robust detection on challenging ETIS images relative to any previous CNN.

Swin-Unet further adopted the shifted window Transformer (Swin-T) into a U-shaped structure~\cite{sota_defeated}. The advanced window aggregation strategy made the attention calculation focused on local patches, but with the connection between different-local-patch windows, so it can be computationally efficient and capable of having global information. Swin-Unet naturally scaled to high-resolution medical images whereas vanilla ViTs not. In practice, this translated to improved localization of the boundaries and increased small polyp detection without practical computation cost. Swin-Unet thus represented a milestone: Transformers could be both accurate and practical at scale in endoscopy.

Polyp-PVT based on the Pyramid Vision Transformer, which can produce multi-scale hierarchical representations like feature pyramids of CNN. This multi-scale property made it possible to model both small-sized and large-sized polyps, one of the limitations in plain ViTs. Via mainly mob pyramid attention architecture, Polyp-PVT gained Dice scores of over 92% on CVC-ClinicDB, and outperformed most CNNs in the field of ETIS with results to show that Transformer models could generalize lesion across lesions scales.

ColonFormer further improved this by creating a polyp specific Transformer encoder. Different from generic ViTs, ColonFormer leveraged domain priors to enhance feature representation of tubular and blob-like structures. Its self-attention mechanism has a good coverage of long-range spatial dependencies, which is useful to capture the context not only from local boundary as well as that from global organs. Offering Dice >93% on CVC-ClinicDB and ~80% on ETIS, it established itself as a new state-of-the-art system especially in terms of generalization.

Lightweight Transformers (LapFormer, Polyp-LVT) mitigated a significant pain point of the Transformer models: giant computational overhead. These models proposed attention pyramids and low-parameter variations of the model that kept much of the global context power, while not hampering its deployment in near real-time. Clinically, this is a trade-off of importance: colonoscopy demands real-time processing (~30 FPS), and lightweight Vits open up an opportunity for practical usage whereas methods based on bulky architecture as in the vanilla Vits do not.

Transformer-based models did well by exploiting self-attention to learn long-range spatial dependencies, which is important due to the fact that subtle polyps might merge into mucosa or emerge from farther distance areas. Architectures such as Swin-Unet and Polyp-PVT further improved robustness through multi-scale representations, while ColonFormer has demonstrated that task-specific adaptation based on medical priors can reinforce clinical relevance. These advances are responsible for their better performance on difficult datasets, such as ETIS. Nevertheless, Transformers encounter some challenges: vanilla ViTs are expensive to compute and may not be suitable for real-time deployment in endoscopy; as a data-hungry model, it is difficult to apply viT into the medical imaging due to annotation constrain; the lightweight version such as Polyp-LVT LapFormer emerge but still struggle with consistent real-time deployment.

Hybrid approaches: Bridging Local and Global

Motivated by the complementary property between CNNs and Transformers, there are a series of hybrid approaches like TransUNet [46], Swin-Unet and their variants proposed in latest works. CNNs are very effective in capturing local, low-level features such as textures, edges and fine boundaries which are highly important for the visualization of polyps. Transformers, on the other hand, model long-range dependencies and global structural patterns that help to provide context information for ambiguous or flat polyp cases which CNNs frequently fail in.

In practice, hybrids combine a CNN front-end encoder for feature extraction and a Transformer backbone to capture spatial dependencies, while the reconstruction process is done following a U-Net like decoder. Detail sensitivity and context awareness are balanced in this architecture. For example, In the TransUNet, CNN-based boundary sharpness has been preserved whilst ViT blocks are utilized to remove false negatives in complex circumstances and seamless Dice scores have been extracted across both popular datasets (Kvasir-SEG, CVC-ClinicDB) and harder benchmarks (ETIS). In the same way, Swin-Unet's windowing mechanism of shifted windows offers hierarchical feature

maps at different scales which increases robustness against varying polyp sizes and balances complexity.

From the clinical point of view, hybrids have been successful as they compensate for CNN's main drawback (little global context), but without giving up deep CFPs' efficiency on medical segmentation problems. Simultaneously, they address Transformers' limitations: hybrids have smaller sample complexity than pure ViTs due to CNN pre-encoding and runs faster than full attention models — important for real-time colonoscopy.

So hybrids are not a mere trade-off -- they represent a clinical sensible evolution by which we see that combining the local feature sensitivity and global attention will generate greater accuracy, efficiency, and generalizability; thus bring computer-aided colonoscopy closer to deployment into routine practice.

Dataset-driven performance differences.

A cross-comparison between the results on Kvasir-SEG, CVC-ClinicDB and ETIS-Larib demonstrates how significantly the dataset characteristics affect segmentation performance. For relatively clean and balanced datasets (e.g., Kvasir-SEG and CVC-ClinicDB) the majority of CNN or transformer-based models obtain superior Dice scores above 85–90% as well. These datasets are packed full of highly visible polyps, good imaging conditions and lesions big enough to be more forgiving to architectural shortcomings.

On the other hand, the ETIS-Larib dataset was proved to be significantly more difficult. Dice scores on CNNs at previous timesteps in a similar manner have dropped here for early CNNs like U-Net and down to 44.9%, with even the ResUNet++ model reaching only 56.2%. There are fewer samples in ETIS, the size and shape of polyps in ETIS are smaller and flatter, and there is more diversity in terms of lighting conditions and imaging quality, which further induces the limitation of local feature extraction that CNN holds. Models that accounted attention and global context explicitly [eg, PraNet (a reverse-attention mechanism) and HarDNet-MSEG architecture with efficient dense connectivity] increased the performance of ETIS performance(64.7% and 73.3%, respectively), but their improvement was not substantial when we compare their performance to those in easier datasets.

Transformerized and hybrid architectures reduced this gap. In capturing long-range dependencies and multi-scale context, TransUNet and Polyp-PVT were also more robust on ETIS and greatly outperformed those CNN-only baselines in terms of Dice and IoU. This indicates that global feature representation is particularly beneficial in challenging clinical environments where polyps have subtle appearance and unclear borders.

In general, dataset-driven discrepancy illustrates two important findings: (1) performance in easy datasets may overestimate the clinical-readiness, as real-world endoscopy typically shares complexity with ETIS; and (2) generalization to challenging datasets serves as an essential benchmark for real deployment, emphasizing both architectural innovation and diverse dataset provision in advancing computer-aided polyp detection and segmentation.

Comparison with previous reviews.

Previous reviews on polyp segmentation studies mainly focused on the CNN-based methods, exhibiting a status of the field till 2021. This study systematically integrates transformer and hybrid approaches, and extends previous work by showing that recent performance advances (since 2022) are largely due to attention. Our findings also reinforce that to some extent (i.e., when benchmark performance approaches saturation on CVC and Kvasir test sets), ETIS is a problem yet to be solved, which constitutes an evidence for the distinction between laboratory benchmarks and real clinical performance.

In conclusion, the comparative study indicates that transformer and hybrid models have become the state of the art in colonoscopic polyps segmentation. However, issues of generalization, limited available datasets and clinical validation are still open challenges. Alleviating these shortcomings will necessitate greater multi-center annotated collections, lightweight model development, and explainable AI architectures promoting clinical trust.

Limitations and Research gaps

While the polyp detection and segmentation has been well developed, there are also some remaining issues. The first reason is that ViTs are more data-hungry than CNNs as only self-attention mechanisms in ViTs scale with the amount of training data. Despite the fact that transfer learning from natural-image datasets, such as ImageNet, gives certain advantages, domain gap between natural and endoscopic images limits performance improvement [34], [35].

Model generalization across clinical centers is also poor, second. Benchmarking frameworks such as EndoCV have shown that many state-of-the-art methods over-fit to individual datasets, and degrade steeply in cross-dataset validation [36]. Current datasets are inadequate to address this issue, which is further intensified due to the limited diversity of current datasets. Despite attempts to mitigate generalizability by initiatives such as multi-center data efforts like PolypGen, dataset bias or efficiency for adding new annotations and the lack of privacy-reasoning are still important challenges [37].

Third, there is a practical limitation due to the computational complexity. With its power of information aggregation, one natural question arises: Can we scale our framework to larger transform based architectures? Large-scale transformer based methods require extensive GPU/TPU resources, which can be expensive to train and difficult in terms of real-time deployment like colonoscopy where inference speeds need to exceed 30 FPS [38], [39]. Light-weight transformers and hybrid models are proposed to alleviate this problem, but balancing accuracy and efficiency is still an issue.

Lastly, explainability and clinical validation still remain significant bottlenecks. Deep models are frequently “black box” in nature, providing little explanation of how decisions were made. This erodes clinical confidence and prevents regulatory approval." There have been calls for interpretable, transparent AI systems following reviews and benchmark studies, but few models that explainability modules can be integrated to show that the effects are still validated in prospective multi-center clinical trials [40], [41], [42].

Trends and Future Outlook

In this context, there are several future research directions for polyp segmentation. CNNs will continue to be powerful baselines for resource-limited scenarios given their efficiency and maturity, and transformer-based approaches will continue to lead state-of-the-art benchmarks thanks in part to their high capacity for global context capture [43], [44].

Hybrids of CNNs and transformers are the future: harnessing the local feature precision from CNNs and long-range dependency modeling of transformers will likely be the way to go. Recent studies have shown that such integrated strategies have systematically achieved better performance than the single models [45], [46].

To alleviate the latter issue of dataset paucity, promising future directions include self-supervised learning, synthetic data generation and domain adaptation methods that endeavor to lower reliance on large amounts of manually labeled data [47]. Efficiency-oriented techniques (e.g., model pruning, quantization and lightweight transformer designs [48]) will facilitate adoption of our method for real-time usage on average colonoscopy equipment.

In addition, explainable AI (XAI) will be essential for clinical uptake. Transparent attention visualization, uncertainty quantification, and clinician-inthe-loop validation should receive more attention [49], [50]. In addition to single-modality imaging, multimodal data (e.g., histopathology, genomics, and patient meta-data) integration is expected to contribute to enhance diagnostic accuracy and robustness leading toward clinically feasible AI-assisted endoscopy [51], [52].

In the end, the next decade will likely see a shift from research prototypes to clinically validated, multimodal and interpretable AI systems that close the divide between academic discovery and real-world clinical utility.

5. Conclusion.

In this work, we have presented a systematic review that captures the development of deep learning methods for colonoscopic polyp detection and segmentation including CNNs, ViTs, and hybrid architectures. Starting from U-Net [5] and evolved into other variants including UNet++ [7],

ResUNet++[6], PraNet [36], the CNN family archit ectures have achieved competitive baseline models with Dice scores between 0.80–0.90 on Kvasir-SEG and CVC-ClinicDB. Nevertheless, their use of local receptive fields hindered generalization and they often performed poorly in challenging datasets like ETIS-Larib (Dice score < 0.60).

ViTs allowed a paradigm shift in the way global dependencies are modelled using self attention, which could help capture more robust subtler features across variations of scale, blurred boundaries and low contrast polyps. Models including Polyp-PVT, ColonFormer and NA-SegFormer outperformed CNNs in all cases with Dice over 0.90 on Kvasir-SEG, ClinicDB and showed reasonably well transferability to ETIS. However, their computational cost and need for data still present challenges for a wide clinical adoption.

Hybrid architectures Hybrid-style architectures, such as TransUNet [14], Swin- UNet [75] and VMDU-Net [8], have arisen as the most promising direction to adapt transformer-based global reasoning with CNN-based local feature extraction. These models proved to be robust and yielded state-of-the-art results, as HarDNet-MSEG and NA-SegFormer managed to surpass 0.70 Dice ETIS threshold that was never crossed by CNN-only anatomical-driven methods.

However, remaining issues have not been addressed such as scarce datasets, computational efficiency and limited clinical validation with different populations. Further progress will also be based on the further development of multimodal datasets, the construction of light and yet accurate architectures, as well as the integration to explainability in our systems to gain clinician trust and regulatory approval.

To summarise, CNNs are still quite efficient baselines, ViTs are the new state-of-the-art and hybrids are a middle-of-the-road route to clinically deployable AI. The fusion of multimodal data and explainable AI frameworks is poised to characterize the next era of computer-aided endoscopy, bringing about earlier diagnosis, enhanced diagnostic yields and ultimately better outcomes in CRC prevention.

Practical Implications

For our clinical partners, these findings indicate that AI-assisted colonoscopy systems are developing fast with hybrid and transformer based models providing improved assistance in diagnosing small or flat polyps often missed in practice." For creating new datasets, the review emphasizes the pressing need for larger, more diverse and multi-center annotated datasets to guarantee generalizability and mitigate bias. For academics, the results highlight trade-offs between lightweight interpretable models and accuracy that are generalisable to real-world performance. Together, these directions are essential for converting algorithmic innovation into clinical practice towards earlier CRC detection and better patient outcomes.

Limitations of This Review

Although the present systematic review attempted to offer an overall summary of deep learning for colonoscopic polyp detection or segmentation, it has several limitations that should be admitted. One, the database search was limited to four databases (PubMed, Scopus, IEEE Xplore and arXiv) and English literature. This may be the reason we neglected potentially relevant studies published in other databases or in non-English journals, which introduced selection bias. Second, even though selection and exclusion criteria met PRISMA recommendations, the final set of studies included in this review may not entirely represent unpublished work, negative results or commercial solutions that are not reported through peer-reviewed channels. Third, the disparate reporting practices between studies (e.g., different dataset splits, preprocessing methodologies and evaluation protocols) affected the direct comparison of performance metrics. For instance, while Dice and IoU were uniformly reported across studies across the board; inference speed (FPS) and computational cost were only obtainable in some of the papers limiting our capability to perform a systematic analysis for real-time feasibility. Fourth, here we mainly focused on image-based models with the datasets like Kvasir-SEG, CVC-ClinicDB, and ETIS-Larib, possibly neglect video-based/multimodal research studies which are becoming more critical for clinical use. Finally, since the field is changing quickly, newly transformer-based or hybrid models developed after early 2025 were not accounted for and their relevance should be evaluated in follow-up updates.

References

[1] R. L. Siegel, N. S. Wagle, A. Cercek, R. A. Smith, and A. Jemal, "Colorectal cancer statistics, 2023," *CA: A Cancer Journal for Clinicians*, vol. 73, no. 3, pp. 233–254, 2023, doi: 10.3322/caac.21772.

[2] J. Bernal, et al., "Comparative validation of polyp detection methods in colonoscopy: Results from the MICCAI 2015 challenge," *Medical Image Analysis*, vol. 31, pp. 1–13, 2016.

[3] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "WM-DOVA Maps for Accurate Polyp Highlighting in Colonoscopy: Validation vs. Saliency Maps from Physicians," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015. (CVC-ClinicDB dataset)

[4] S. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 9, no. 2, pp. 283–293, 2014. (ETIS-Larib Polyp DB)

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

[6] D. Jha, M. Smedsrud, M. A. Riegler, D. Johansen, T. de Lange, P. Halvorsen, and H. D. Johansen, "ResUNet++: An advanced architecture for medical image segmentation," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, 2019, pp. 225–2255.

[7] Z. Zhou, M. M. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (MICCAI Workshop)*, 2018, pp. 3–11.

[8] G. Urban, P. Tripathi, T. Alkayali, M. Mittal, F. Jalali, W. Karnes, and P. Baldi, "Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy," *Gastroenterology*, vol. 155, no. 4, pp. 1069–1078, 2018.

[9] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. ECCV*, 2018, pp. 801–818.

[10] O. Urban, et al., "Deep learning for real-time detection of colorectal polyps in colonoscopy videos," *The Lancet Oncology*, vol. 19, no. 7, pp. 793–800, 2018.

[11] D. Jha, P. Halvorsen, H. D. Johansen, D. Johansen, T. de Lange, and M. A. Riegler, "Kvasir-SEG: A segmented polyp dataset," in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, 2020, pp. 4050–4054.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An image is worth 16×16 words: Transformers for image recognition at scale," *arXiv preprint* arXiv:2010.11929, 2020.

[13] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, D. Johansen, T. de Lange, and H. D. Johansen, "ResUNet++: An advanced architecture for medical image segmentation," in *Proc. IEEE Int. Symp. Biomed. Imaging (ISBI)*, 2019, pp. 223–227.

[14] D. Jha, S. Ali, N. K. Tomar, H. D. Johansen, D. Johansen, M. A. Riegler, P. Halvorsen, and T. de Lange, "ColonSegNet: A dilated convolutional neural network for colon polyp segmentation," in *Proc. IEEE Int. Symp. Computer-Based Medical Systems (CBMS)*, 2021, pp. 191–196.

[15] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?," in *Proc. NeurIPS*, 2021.

[16] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-Unet: Unet-like pure transformer for medical image segmentation," in *Proc. MICCAI*, 2021, pp. 100–110.

[17] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," *arXiv preprint* arXiv:2102.04306, 2021.

[18] H. Huang, L. Lin, R. Tong, and G. Hu, "HarDNet-MSEG: A low memory requirement network for polyp segmentation," *Medical Image Analysis*, vol. 75, p. 102304, 2022.

[19] S. Ali, H. Realdon, et al., "An objective comparison of polyp detection methods in colonoscopy: EndoCV 2021 challenge," *Medical Image Analysis*, vol. 77, p. 102336, 2022.

[20] R. Chen, et al., "Semi-supervised learning methods for polyp segmentation: A review," *Medical Image Analysis*, vol. 82, p. 102639, 2022.

[21] M.-H. Guo, C. Xu, J. Liu, et al., "Polyp-PVT: Polyp segmentation with pyramid vision transformers," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 6, pp. 3120–3130, 2022.
[22] Z. Dong, H. Cao, et al., "ColonFormer: Effective transformer-based polyp segmentation," *Medical Image Analysis*, vol. 86, p. 102792, 2023.
[23] VCIBA Consortium, "Transformers for medical image analysis: A tutorial review," *Visual Computing for Industry, Biomedicine, and Art*, vol. 6, no. 1, 2023, doi: 10.1186/s42492-023-00138-6.
[24] R. Wang, Y. Zhang, X. Li, and J. Sun, "NA-SegFormer: Neural architecture search for transformer-based polyp segmentation," *IEEE Transactions on Medical Imaging*, vol. 42, no. 1, pp. 15–27, 2023.
[25] J. Lee, M. Kim, H. Park, and S. Lim, "Polyp-LVT: Lightweight vision transformer for efficient polyp segmentation," *Medical Image Analysis*, vol. 91, p. 103025, 2024.
[26] S. Kumar, P. Rani, and R. Gupta, "Tiny polyp detection using lightweight CNNs," *Pattern Analysis and Applications*, 2024. (?? not yet verified).
[27] S. Ali, J. Jha, M. Smedsrud, D. Johansen, P. Halvorsen, H. D. Johansen, et al., "PolypGen: A multi-center polyp detection and segmentation dataset for generalisability assessment," *arXiv preprint* arXiv:2106.04463, 2021.
[28] Y. Oukdach, A. Garbaz, Z. Kerkaou, M. El Ansari, L. Koutti, A. F. El Ouafdi, and M. Salihoun, "UViT-Seg: An efficient ViT and U-Net-based framework for accurate colorectal polyp segmentation in colonoscopy and WCE images," *Journal of Digital Imaging*, vol. 37, no. 5, pp. 2354–2374, 2024, doi: 10.1007/s10278-024-01124-8.
[29] Q. H. Trinh, N. T. Bui, T. H. Nguyen Mau, M. V. Nguyen, H. M. Phan, M. T. Tran, and H. D. Nguyen, "M^2UNet: MetaFormer multi-scale upsampling network for polyp segmentation," *arXiv preprint* arXiv:2306.08600, 2023, doi: 10.48550/arXiv.2306.08600.
[30] L. Wang, Z. Liu, and Q. Li, "Nested UNet++ for colonoscopy segmentation," *Medical Physics*, 2024. (?? not yet verified).
[31] T. Li, P. Liu, and H. Zhang, "Real-time polyp detection with lightweight transformers," *IEEE Transactions on Medical Imaging*, 2024.
[32] Y. Zhao, L. Huang, and F. Wang, "LapFormer: Lightweight attention pyramid transformers for polyp segmentation," *arXiv preprint* arXiv:2210.04393, 2024. (?? not yet verified).
[33] R. Chen, X. Ma, and J. Zhang, "NA-SegFormer with neighborhood attention achieving 96% accuracy," *Scientific Reports*, vol. 14, pp. 1–12, 2024.
[34] Y. Khan, B. Ahmed, and A. Raza, "Hybrid CNN–Transformer methods for polyp segmentation," *arXiv preprint* arXiv:2508.09189, 2025. (?? future work).
[35] P. Lijin, M. Ullah, A. Vats, F. A. Cheikh, G. S. Kumar, and M. S. Nair, "PolySegNet: Improving polyp segmentation through Swin Transformer and Vision Transformer fusion," *Biomedical Engineering Letters*, vol. 14, pp. 1421–1431, Aug. 2024, doi: 10.1007/s13534-024-00355-7.
[36] D. Fan, G. Ji, T. Zhou, G. Chen, H. Fu, and L. Shao, "PraNet: Parallel reverse attention network for polyp segmentation," *Medical Image Analysis*, vol. 72, p. 102084, 2021.
[37] S. Srivastava, P. Jha, and A. Jha, "MSRF-Net: A multi-scale residual fusion network for biomedical image segmentation," *Computers in Biology and Medicine*, vol. 134, p. 104427, 2021.
[38] T. Huang, Y. Xu, Y. Song, X. Yan, Y. Zhang, and Y. Wang, "SSFormer: A lightweight structure-shared transformer for medical image segmentation," *Medical Image Analysis*, vol. 84, p. 102684, 2023.
[39] S. Ding, J. Zhang, and J. Yu, "Feature cross-bridging transformer for medical image segmentation," *Knowledge-Based Systems*, vol. 269, p. 110481, 2023.
[40] X. Yang, Z. Zhang, and L. Zhu, "A lighter hybrid feature fusion framework for polyp segmentation," *Biomedical Signal Processing and Control*, vol. 97, p. 105735, 2024.
[41] W. Li, Y. Chen, J. Wang, and H. Xu, "VMDU-Net: Vision Mamba Dual-encoder UNet for accurate polyp segmentation," *Knowledge-Based Systems*, vol. 311, p. 111999, 2025.
[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 6000–6010.
[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[44] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. CVPR*, 2015, pp. 1–9.
[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint* arXiv:1409.1556, 2014.
[46] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NeurIPS*, 2012, pp. 1097–1105.
[47] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, 2009, pp. 248–255.
[48] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.
[49] T. Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. ECCV*, 2014, pp. 740–755.
[50] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3523–3542, 2022.
[51] J. Mei, T. Zhou, K. Huang, Y. Zhang, Y. Zhou, Y. Wu, and H. Fu, "A survey on deep learning for polyp segmentation: Techniques, challenges and future trends," *arXiv preprint* arXiv:2311.18373, 2023.
[52] Z. Wu, F. Lv, C. Chen, A. Hao, and S. Li, "Colorectal polyp segmentation in the deep learning era: A comprehensive survey," *arXiv preprint* arXiv:2401.11734, 2024.