Computers, Materials &
Continua

Tech Science Press

Doi:10.32604/cmc.2025.074416

ARTICLE

HMA-DER: A Hierarchical Attention and Expert Routing Framework for Accurate Gastrointestinal Disease Diagnosis

Sara Tehsin[1], **Inzamam Mashood Nasir**[1,*], **Wiem Abdelbaki**[2], **Fadwa Alrowais**[3], **Khalid A. Alattas**[4], **Sultan Almutairi**[5] **and Radwa Marzouk**[6]

[1]Faculty of Informatics, Kaunas University of Technology, Kaunas, 51368, Lithuania
[2]College of Engineering and Technology, American University of the Middle East, Egaila, 54200, Kuwait
[3]Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh, 11671, Saudi Arabia
[4]Department of Computer Science and Artificial Intelligence, College of Computer Science and Engineering, University of Jeddah, Jeddah, 23890, Saudi Arabia
[5]Department of Computer Science, Applied College, Shaqra University, Shaqra, 15526, Saudi Arabia
[6]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh, 11671, Saudi Arabia

*Corresponding Author: Inzamam Mashood Nasir. Email: inzamam.nasir@ktu.edu

Received: 10 October 2025; Accepted: 17 November 2025; Published: 10 February 2026

ABSTRACT: Objective: Deep learning is employed increasingly in Gastroenterology (GI) endoscopy computer-aided diagnostics for polyp segmentation and multi-class disease detection. In the real world, implementation requires high accuracy, therapeutically relevant explanations, strong calibration, domain generalization, and efficiency. Current Convolutional Neural Network (CNN) and transformer models compromise border precision and global context, generate attention maps that fail to align with expert reasoning, deteriorate during cross-center changes, and exhibit inadequate calibration, hence diminishing clinical trust. **Methods:** HMA-DER is a hierarchical multi-attention architecture that uses dilation-enhanced residual blocks and an explainability-aware Cognitive Alignment Score (CAS) regularizer to directly align attribution maps with reasoning signals from experts. The framework has additions that make it more resilient and a way to test for accuracy, macro-averaged F1 score, Area Under the Receiver Operating Characteristic Curve (AUROC), calibration (Expected Calibration Error (ECE), Brier Score), explainability (CAS, insertion/deletion AUC), cross-dataset transfer, and throughput. **Results:** HMA-DER gets Dice Similarity Coefficient scores of 89.5% and 86.0% on Kvasir-SEG and CVC-ClinicDB, beating the strongest baseline by +1.9 and +1.7 points. It gets 86.4% and 85.3% macro-F1 and 94.0% and 93.4% AUROC on HyperKvasir and GastroVision, which is better than the baseline by +1.4/+1.6 macro-F1 and +1.2/+1.1 AUROC. Ablation study shows that hierarchical attention gives the highest (+3.0), followed by CAS regularization (+2–3), dilatation (+1.5–2.0), and residual connections (+2–3). Cross-dataset validation demonstrates competitive zero-shot transfer (e.g., KS→CVC Dice 82.7%), whereas multi-dataset training diminishes the domain gap, yielding an 88.1% primary-metric average. HMA-DER's mixed-precision inference can handle 155 pictures per second, which helps with calibration. **Conclusion:** HMA-DER strikes a compromise between accuracy, explainability, robustness, and efficiency for the use of reliable GI computer-aided diagnosis in real-world clinical settings.

KEYWORDS: Gastrointestinal image analysis; polyp segmentation; multi-attention deep learning; explainable AI; cognitive alignment score; cross-dataset generalization

1 Introduction

Deep learning has improved computer-aided diagnosis (CAD) for gastrointestinal (GI) endoscopy, allowing for polyp segmentation and the detection of many disease classes in different clinical settings. There are a number of public datasets, including Kvasir-SEG (KS) for pixel-level polyp delineation [1], CVC-ClinicDB (CVC) for benchmark segmentation in difficult imaging conditions [2,3], HyperKvasir (HK) for large-scale, heterogeneous classification [4], and GastroVision (GV) for curated multi-class diagnosis [5]. Even while accuracy has gotten better, routine clinical use still needs calibrated probability, robustness to domain shifts, and explanations that are in line with what humans would expect to generate trust in the system.

Residual and densely linked networks have enhanced optimization and feature reuse in medical imaging, creating robust baselines [6,7]. Newer vision transformers, such as ViT and Swin, have shown higher performance thanks to global self-attention. Hybrid designs, like TransUNet [8] and MedT, combine convolutional priors with transformer modules to better model context in medical pictures [8,9]. Three things make it hard to use GI CAD: (i) getting accurate estimates of fine lesion borders and long-range context without overfitting; (ii) giving calibrated confidence estimates for making decisions based on risk; and (iii) keeping accuracy when cross-center distribution shifts and common corruptions happen.

To highlight the novelty and significance of this work, the key contributions of the proposed HMA-DER framework are summarized as follows:

- We introduce HMA-DER, a unified framework that combines Hierarchical Multi-Attention (HMA) with Dynamic Expert Routing (DER) to achieve accurate and explainable gastrointestinal (GI) disease diagnosis from endoscopic images.
- A hierarchical multi-attention system collects information that is useful at the global, regional, and local levels. This helps the model focus on clinically important features, including polyps, ulcer boundaries, bleeding, and irritated mucosa, while getting rid of background noise.
- Dynamic expert routing sends low-confidence or complicated samples to specialized expert networks to make classification more reliable and cut down on diagnostic mistakes like false negatives and positives.
- The integrated explainability module inherently performs lesion localization during classification, producing interpretable attention maps that align closely with expert-annotated lesion regions.
- The system for segmentation and multi-class classification is tested using four complimentary benchmark datasets: Kvasir-SEG, CVC-ClinicDB, HyperKvasir, and GastroVision. HMA-DER often beats advanced CNN and Transformer baselines.
- Ablation experiments demonstrate the efficacy of the hierarchical attention and expert routing modules, while clinical alignment metrics indicate that the model's attention correlates with actual pathological structures.
- The proposed system establishes a practical step toward clinically trustworthy computer-aided diagnosis by combining high diagnostic accuracy with interpretable, anatomically grounded visual reasoning.

The structure of this paper continues below. Section 2 looks at deep learning research on diagnosing gastrointestinal diseases, focusing on attention mechanisms and expert systems. In Section 3, we talk about the HMA-DER framework's hierarchical multi-attention architecture and dynamic expert routing method. Section 4 talks about the datasets, implementation settings, evaluation criteria, and experimental findings. These include ablation experiments and clinical alignment studies. In Section 5, the study ends with important results and ideas for further research.

2 Related Work

Computer-aided diagnosis (CAD) for gastrointestinal endoscopy has progressed swiftly owing to deep learning, with public datasets for reproducible benchmarking in segmentation and classification tasks. Kvasir-SEG and CVC-ClinicDB standardize polyp delineation at the pixel level, whereas HyperKvasir and GastroVision provide extensive, diverse, multi-class image cohorts for diagnostic recognition [4,5]. These corpora have sped up the shift from handwritten features to learning based on convolutional and transformer networks from start to finish. FCN and U-Net were used in the early days of medical picture segmentation to create encoder-decoder principles and skip connections for precise localization [10,11]. Later improvements, such as PSPNet and DeepLabv3+, included pyramid pooling, atrous convolution, and encoder-decoder tuning to improve global context capture [12]. In medical domains, nnU-Net demonstrated that careful data- and task-adaptive configuration can rival bespoke architectures across diverse modalities [13]. For colonoscopic polyp segmentation specifically, reverse-attention and boundary-aware designs (e.g., PraNet) pushed the state of the art by suppressing distractors and refining edges [14]. These CNN-centric trends established strong baselines but can struggle to reconcile fine boundary detail with long-range global context under acquisition variability.

Transformer models introduced global self-attention to medical vision, narrowing the gap to or surpassing CNNs in both classification and dense prediction. Vision Transformers and hierarchical variants such as Shifted Window Transformer capture long-range dependencies with scalable receptive fields [15,16]. Hybrid and U-shaped transformers tailored to segmentation (e.g., TransUNet, MedT, and Swin-Unet) combine convolutional priors with token mixing to balance local detail and global semantics [8,9,17]. Despite these gains, attention distributions are not guaranteed to align with clinically relevant cues, and training deep attention stacks can be sensitive to optimization and data shift.

For GI CAD to be reliable, it needs to be able to explain itself. Gradient-based variations and class activation mapping pinpoint prediction evidence, whereas axiomatic attributions such as Integrated Gradients amplify feature contribution sensitivity [18]. Saliency methods may not pass sanity checks, necessitating more stringent evaluation and training time limitations [19]. Using insertion/deletion curves, perturbation-based audits such as meaningful perturbations and randomized input sampling for explanation (RISE) assess causal fidelity [20,21]. "Right-for-the-right-reasons" objectives use explanation targets to limit models, which helps people think in a way that is consistent with their values. This goes beyond just looking at the results after the fact. This study proposes the integration of architectural attention methodologies with direct supervision that directs explanations towards clinically pertinent areas.

HMA-DER brings together multi-scale representation learning, explanation alignment, and strong training. The technique uses U-shaped encoder-decoder principles, global context, and dilation in residual blocks to make optimization more stable. It also has an explanation-aware regularizer that uses "right-for-the-right-reasons" training and perturbation-faithfulness assessments [20,21]. The method tackles accuracy, interpretability, calibration, robustness, and throughput for clinically viable GI CAD within a cohesive pipeline assessed on KS, CVC, HK, and GV [1–5].

3 Proposed Methodology

The Hierarchical Multi-Stage Attention with Dynamic Expert Routing (HMA-DER) architecture correctly and understandably diagnoses gastrointestinal (GI) illnesses. The framework has hierarchical feature extraction for learning multi-resolution representations, stage-wise attention mechanisms that narrow the focus from global anatomy to localized lesions and micro-patterns, and a dynamic expert routing module that clearly gives decision-making power to specialized sub-networks. These characteristics render diagnostic predictions performance-oriented and therapeutically relevant. The following subsections present the methodology in detail, accompanied by formal mathematical formulations that rigorously define each stage of the proposed system. As shown in Fig. 1, HMA-DER proceeds from preprocessing to hierarchical features, three attention stages, dynamic expert routing, and explainable inference.

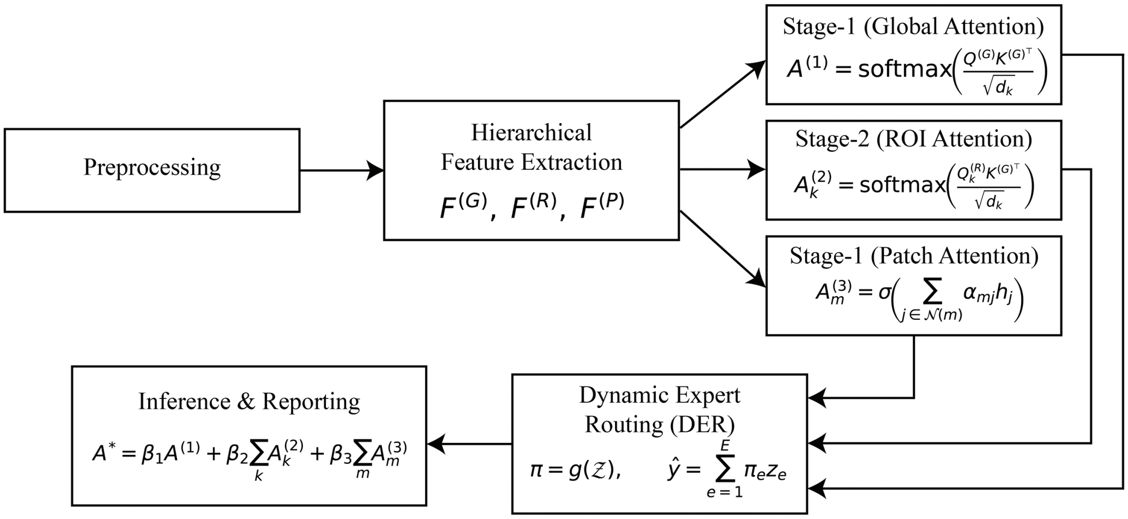
Preprocessing
Hierarchical Feature Extraction
$F^{(G)}, F^{(R)}, F^{(P)}$
Stage-1 (Global Attention)
$A^{(1)} = \mathrm{softmax}\left(\frac{Q^{(G)}K^{(G)\top}}{\sqrt{d_k}}\right)$
Stage-2 (ROI Attention)
$A_k^{(2)} = \mathrm{softmax}\left(\frac{Q_k^{(R)}K^{(G)\top}}{\sqrt{d_k}}\right)$
Stage-1 (Patch Attention)
$A_m^{(3)} = \sigma\left(\sum_{j\in\mathcal{N}(m)} \alpha_{mj}h_j\right)$
Dynamic Expert Routing (DER)
$\pi = g(\mathcal{Z}), \quad \hat{y} = \sum_{e=1}^{E}\pi_e z_e$
Inference & Reporting
$A^* = \beta_1 A^{(1)} + \beta_2\sum_k A_k^{(2)} + \beta_3\sum_m A_m^{(3)}$

Figure 1: HMA-DER pipeline overview. The method flows from preprocessing to hierarchical feature extraction, Stage-1: global attention, Stage-2: ROI attention, Stage-3: patch attention, dynamic expert routing, and explainable inference with fused heatmaps and textual rationale. Notation: $A^{(1)}$, $A_k^{(2)}$, $A_m^{(3)}$ are attention maps; π are routing weights; $\hat{y}$ is the final prediction

3.1 Input Representation and Preprocessing

Let the training corpus be denoted by $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where N represents the total number of samples. Each element $x_i \in \mathbb{R}^{H\times W\times 3}$ corresponds to a color endoscopic frame of spatial resolution $H \times W$ with three channels representing the red, green, and blue intensities. The associated label $y_i \in \{1, 2, \ldots, C\}$ assigns the image to one of the C gastrointestinal disease classes, such as polyps, ulcers, bleeding, or normal tissue.

Since raw endoscopic frames are often affected by variable illumination, device-specific contrast, and motion-induced artifacts, a preprocessing operator $\phi(\cdot)$ is applied to every input image. The preprocessing is designed to normalize the pixel intensities while ensuring that diagnostically relevant regions are not distorted. Mathematically, this operator can be expressed as

$$\tilde{x}_i = \phi(x_i) = \frac{x_i - \mu}{\sigma}, \tag{1}$$

where $\tilde{x}_i$ is the standardized image after preprocessing. The parameter $\mu \in \mathbb{R}^3$ denotes the channel-wise mean vector, computed across the dataset, such that $\mu = [\mu_R, \mu_G, \mu_B]^\top$. Similarly, $\sigma \in \mathbb{R}^3$ represents the channel-wise standard deviation vector $\sigma = [\sigma_R, \sigma_G, \sigma_B]^\top$. Subtracting the mean μ from the raw image x_i centers the pixel intensity distribution around zero, while dividing by the standard deviation σ scales the distribution to unit variance. This transformation yields images $\tilde{x}_i$ that are statistically normalized and more amenable to stable training of deep feature extractors.

In addition to normalization, domain-specific corrections are employed. Illumination in gastrointestinal imaging is highly non-uniform due to the narrow-beam light sources in endoscopes. To correct for this, an illumination equalization function $\psi(\cdot)$ is applied such that

$$\hat{x}_i = \psi(\tilde{x}_i) = \frac{\tilde{x}_i}{\alpha + \|\tilde{x}_i\|_p}, \tag{2}$$

where $\alpha > 0$ is a small stabilizing constant and $\|\tilde{x}_i\|_p$ denotes the p-norm of the intensity distribution. This operation ensures that excessively bright or dark regions are compressed into a consistent dynamic range. Finally, to remove specular highlights and border artifacts without altering pathological content, a masking operator $\mathcal{M}(\cdot)$ is applied:

$$x_i^* = \mathcal{M}(\hat{x}_i) = \hat{x}_i \odot m_i, \tag{3}$$

where $\odot$ indicates element-wise multiplication and $m_i \in \{0, 1\}^{H\times W}$ is a binary mask that suppresses non-informative pixels (such as black borders or instrument reflections). The resulting preprocessed image x_i^* is thus normalized, illumination-corrected, and artifact-suppressed, ensuring that subsequent attention mechanisms highlight true diagnostic structures.

The approach changes raw endoscopic pictures x_i into a standardized representation x_i^* that is stable for deep learning, resistant to changes in acquisition, and reliably shows clinically important aspects. Preprocessing does not add fake patterns that could make downstream attention maps hard to understand; therefore, this careful design helps explainability.

3.2 Hierarchical Feature Extraction

After processing, images x_i^* are sent to a hierarchical backbone network $\mathcal{F}(\cdot)$. CNNs (convolutional neural networks) pick up on local spatial textures, whereas transformer blocks model long-range relationships and interactions in context. This hybrid architecture takes the best parts of each. The integrated architecture keeps learnt representations of fine-grained lesion patterns and the whole body.

Formally, let $\tilde{x} \in \mathbb{R}^{H\times W\times 3}$ denote a preprocessed image. The hierarchical backbone decomposes $\tilde{x}$ into three progressively refined levels of feature maps:

$$F^{(G)} = \mathcal{F}_G(\tilde{x}), \qquad F^{(R)} = \mathcal{F}_R(\tilde{x}), \qquad F^{(P)} = \mathcal{F}_P(\tilde{x}), \tag{4}$$

where $F^{(G)}$, $F^{(R)}$, and $F^{(P)}$ represent the global, region-of-interest (ROI), and patch-level feature spaces, respectively. Fig. 2 depicts the three feature tiers that support explanations across organ, lesion, and micro-pattern scales.

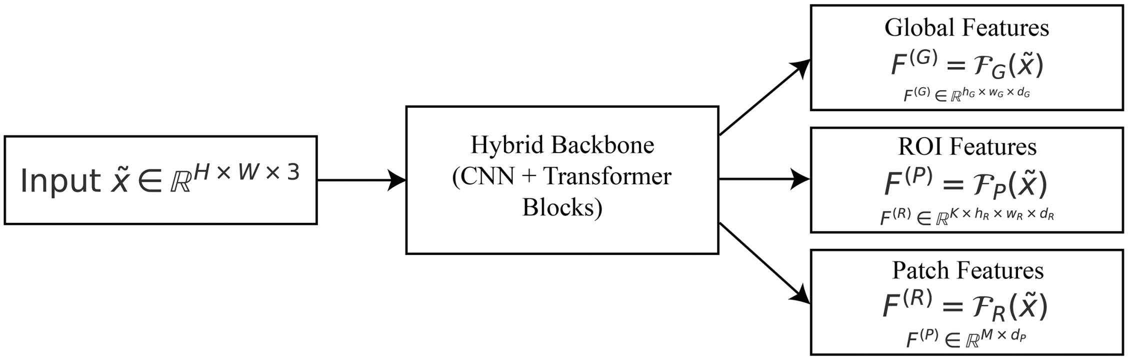
Input $\tilde{x} \in \mathbb{R}^{H\times W\times 3}$
Hybrid Backbone (CNN + Transformer Blocks)
Global Features
$F^{(G)} = \mathcal{F}_G(\tilde{x})$
$F^{(G)} \in \mathbb{R}^{h_G\times w_G\times d_G}$
ROI Features
$F^{(P)} = \mathcal{F}_P(\tilde{x})$
$F^{(R)} \in \mathbb{R}^{K\times h_R\times w_R\times d_R}$
Patch Features
$F^{(R)} = \mathcal{F}_R(\tilde{x})$
$F^{(P)} \in \mathbb{R}^{M\times d_P}$

Figure 2: Hierarchical features: $F^{(G)} = \mathcal{F}_G(\tilde{x})$ (global organ context), $F^{(R)} = \mathcal{F}_R(\tilde{x})$ (ROI/lesion descriptors), and $F^{(P)} = \mathcal{F}_P(\tilde{x})$ (fine patch embeddings)

To get contextual information from the whole image, deep convolutional filters and a transformer encoder are used to make the global feature representation $F^{(G)} \in \mathbb{R}^{h_G \times w_G \times d_G}$. After downsampling, h_G and w_G show the decreased spatial resolution, while d_G shows the feature embedding dimension. These features depict organ walls, lumen boundaries, and coarse disease indications.

The region-level feature representation $F^{(R)} \in \mathbb{R}^{K \times h_R \times w_R \times d_R}$ is generated by choosing K candidate regions of interest from the global feature map. ROIs are adaptively pooled to a fixed spatial resolution of $h_R \times w_R$ and then encoded into a d_R-dimensional feature space. The collection $\{F_k^{(R)}\}_{k=1}^{K}$ focuses on lesion-centric information since k indexes each ROI. This step links the global structural background to the description of local abnormalities.

The patch-level feature representation $F^{(P)} \in \mathbb{R}^{M \times d_P}$ aggregates detailed micro-patterns from small patches $\{P_m\}_{m=1}^{M}$, situated either near suspected lesions or uniformly across the image. A transformer block with self-attention encodes each patch $P_m \in \mathbb{R}^{p \times p \times 3}$ into a d_P-dimensional embedding to mimic spatial interactions among patches. These features find small diagnostic signals that coarser feature maps overlook, such as vascular anomalies, pit patterns, and ulcer borders.

In summary, the mapping in (4) provides a multi-resolution hierarchy:

$$\tilde{x} \xrightarrow{\mathcal{F}_G} F^{(G)} \quad \rightarrow \quad \mathcal{F}_R(F^{(G)}) = F^{(R)} \quad \rightarrow \quad \mathcal{F}_P(F^{(R)}) = F^{(P)}. \tag{5}$$

Global descriptors keep the context at the organ level, ROI descriptors find lesion candidates, and patch-level descriptors figure out microstructures that are different. This hierarchical breakdown lets the attention modules downstream explain the whole gastrointestinal tract down to the smallest odd patterns.

3.3 Stage-1: Global Attention (Coarse Localization)

After getting hierarchical feature maps, global reasoning is employed to discover the anatomical zones that are most helpful for diagnosis. This level roughly places the model toward the lumen boundaries, mucosal areas, and big problems.

Let $F^{(G)} \in \mathbb{R}^{h_G \times w_G \times d_G}$ denote the global feature representation obtained from the backbone. To apply attention, these features are linearly projected into three distinct spaces: queries, keys, and values. This is expressed as

$$Q^{(G)} = F^{(G)} W_Q, \qquad K^{(G)} = F^{(G)} W_K, \qquad V^{(G)} = F^{(G)} W_V, \tag{6}$$

where W_Q, W_K, $W_V \in \mathbb{R}^{d_G \times d_k}$ are learnable projection matrices that map the d_G-dimensional embeddings into a common latent space of dimension d_k. The query matrix $Q^{(G)}$ encodes what the model is “looking for,” the key matrix $K^{(G)}$ encodes where this information resides in the global feature map, and the value matrix $V^{(G)}$ provides the actual content to be aggregated. The attention scores are computed by measuring the similarity between queries and keys:

$$A^{(1)} = \mathrm{softmax}\left(\frac{Q^{(G)} K^{(G)\top}}{\sqrt{d_k}}\right), \tag{7}$$

where the dot product $Q^{(G)} K^{(G)\top}$ evaluates alignment between query and key vectors. Global attention and the resulting $Z^{(1)}$ are visualized in Fig. 3.

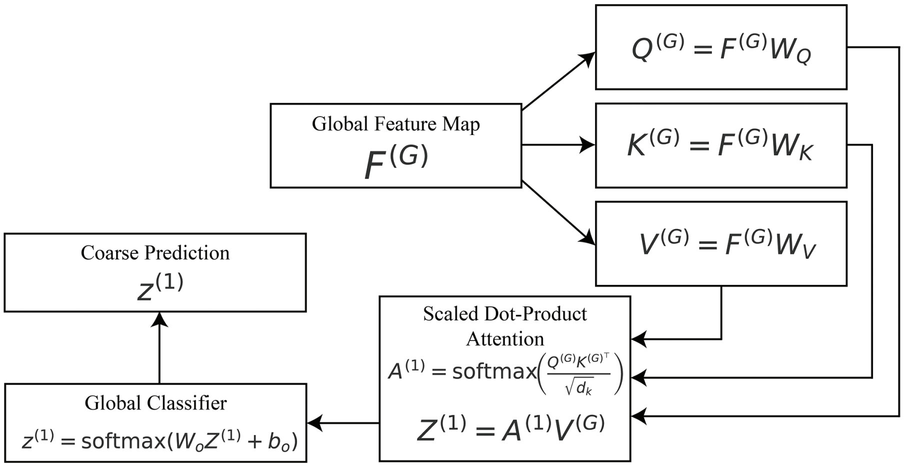
Global Feature Map
$F^{(G)}$
$Q^{(G)} = F^{(G)}W_Q$
$K^{(G)} = F^{(G)}W_K$
$V^{(G)} = F^{(G)}W_V$
Scaled Dot-Product Attention
$A^{(1)} = \mathrm{softmax}\left(\frac{Q^{(G)}K^{(G)\top}}{\sqrt{d_k}}\right)$
$Z^{(1)} = A^{(1)}V^{(G)}$
Global Classifier
$z^{(1)} = \mathrm{softmax}(W_oZ^{(1)} + b_o)$
Coarse Prediction
$z^{(1)}$

Figure 3: Stage-1 global attention: $A^{(1)} = \mathrm{softmax}\big(Q^{(G)}K^{(G)\top}/\sqrt{d_k}\big)$, $Z^{(1)} = A^{(1)}V^{(G)}$, and $z^{(1)} = \mathrm{softmax}(W_oZ^{(1)} + b_o)$. Coarse saliency localizes organ regions relevant to diagnosis

The division by $\sqrt{d_k}$ ensures numerical stability by preventing the inner products from becoming excessively large as the dimensionality increases. The softmax$(\cdot)$ function normalizes each row of the similarity matrix so that the attention weights are positive and sum to one, thereby forming a probability distribution over spatial locations. Using these attention weights, the attended representation is formed as

$$Z^{(1)} = A^{(1)}V^{(G)}, \tag{8}$$

where $Z^{(1)} \in \mathbb{R}^{h_G \times w_G \times d_k}$ represents the reweighted global features that emphasize diagnostically important regions while suppressing irrelevant background. In clinical terms, $Z^{(1)}$ highlights organ-level structures such as the colon wall or gastric folds, allowing the model to direct subsequent analysis toward medically meaningful zones. Finally, a classification head produces coarse global predictions:

$$z^{(1)} = \mathrm{softmax}(W_oZ^{(1)} + b_o). \tag{9}$$

$W_o \in \mathbb{R}^{d_k \times C}$ and $b_o \in R^C$ are parameters that can be learned and that project the attended representation onto the C-dimensional class space. The vector $z^{(1)} \in \mathbb{R}^C$ contains the class probabilities from the model's first diagnosis, which was based exclusively on global anatomical inputs.

To show why the model made its choice, plot the attention matrix $A^{(1)}$ as a heatmap over the original image. This will show which big portions of the gastrointestinal system are important. Clinicians can trust a more detailed lesion analysis by analyzing these maps to see if the model is focusing on the right anatomical structures.

3.4 Stage-2: ROI Attention (Lesion-Aware Refinement)

After getting global attention and feature representation in Stage 1, the framework moves on to potential lesions for a more in-depth look. This is because gastrointestinal issues, including polyps, ulcers, and bleeding areas, are usually limited; therefore, a wide description could miss their diagnostic details. A lightweight lesion proposal module $\mathcal{P}(\cdot)$ generates candidate regions of interest (ROIs) from preprocessed images or global feature maps.

$$\{R_k\}_{k=1}^{K} = \mathcal{P}(\tilde{x}), \tag{10}$$

where K denotes the total number of proposed regions, and $R_k \in \mathbb{R}^{h_R \times w_R \times 3}$ represents the k-th cropped patch aligned with a potential lesion area.

Each ROI R_k is embedded into a feature representation $F_k^{(R)} \in \mathbb{R}^{h'_R \times w'_R \times d_R}$ using convolutional and transformer layers, analogous to the processing in the global stage but with localized context. Queries derived from these ROI features are denoted as

$$Q_k^{(R)} = F_k^{(R)} W_Q^{(R)}, \qquad W_Q^{(R)} \in \mathbb{R}^{d_R \times d_k}, \tag{11}$$

where $W_Q^{(R)}$ is a learnable projection matrix. The keys and values are inherited from the global representation, namely $K^{(G)}$ and $V^{(G)}$, so that the region-level analysis remains grounded in the global context of the organ. The cross-attention between an ROI and the global feature map is computed as

$$A_k^{(2)} = \text{softmax}\left(\frac{Q_k^{(R)} K^{(G)\top}}{\sqrt{d_k}}\right), \tag{12}$$

where $A_k^{(2)} \in \mathbb{R}^{(h'_R w'_R) \times (h_G w_G)}$ represents the attention weights linking the spatial locations of the k-th ROI to all spatial positions of the global feature map. The normalization ensures that the model assigns interpretable probabilities to global locations in relation to each ROI. Fig. 4 shows how ROI queries attend to global context to refine lesion evidence.

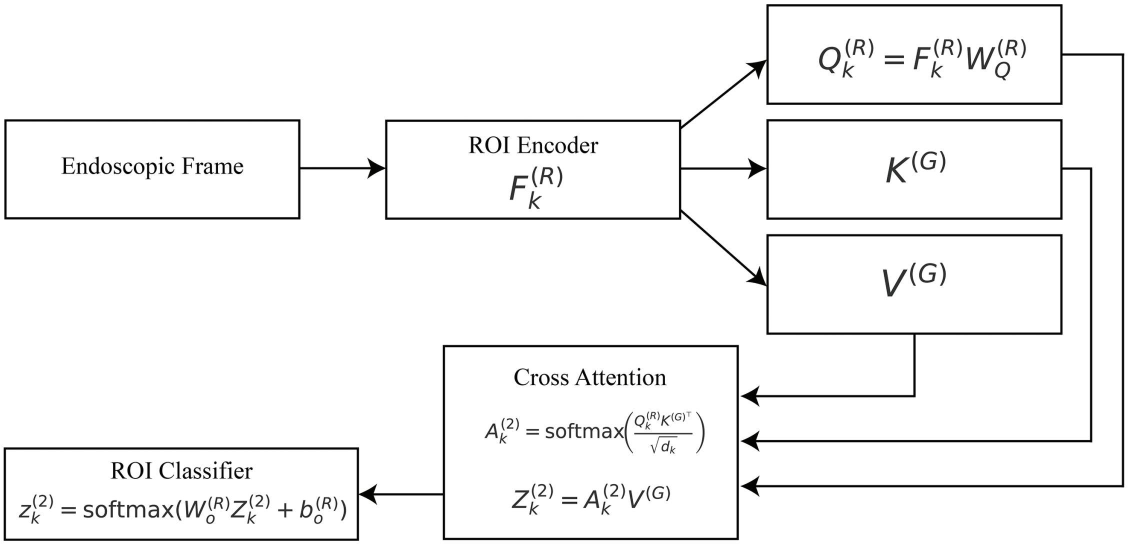
Endoscopic Frame
ROI Encoder
$F_k^{(R)}$
$Q_k^{(R)} = F_k^{(R)} W_Q^{(R)}$
$K^{(G)}$
$V^{(G)}$
Cross Attention
$A_k^{(2)} = \text{softmax}\left(\frac{Q_k^{(R)} K^{(G)\top}}{\sqrt{d_k}}\right)$
$Z_k^{(2)} = A_k^{(2)} V^{(G)}$
ROI Classifier
$z_k^{(2)} = \text{softmax}(W_o^{(R)} Z_k^{(2)} + b_o^{(R)})$

Figure 4: Stage-2 ROI attention: $A_k^{(2)} = \text{softmax}\left(Q_k^{(R)} K^{(G)\top} / \sqrt{d_k}\right)$ and $Z_k^{(2)} = A_k^{(2)} V^{(G)}$ yield lesion-focused descriptors and logits $z_k^{(2)}$

The attended representation of the k-th ROI is then

$$Z_k^{(2)} = A_k^{(2)} V^{(G)}, \tag{13}$$

where $Z_k^{(2)} \in \mathbb{R}^{h'_R \times w'_R \times d_k}$ encodes features that fuse local ROI information with relevant global anatomical context. This step enriches the lesion representation by embedding it into the broader tissue environment, reflecting the clinical intuition that the significance of a lesion often depends on its surrounding mucosal structures.

Finally, a classification head is applied to each ROI to generate lesion-specific logits:

$$z_k^{(2)} = \text{softmax}(W_o^{(R)} Z_k^{(2)} + b_o^{(R)}), \tag{14}$$

where $W_o^{(R)} \in \mathbb{R}^{d_k \times C}$ and $b_o^{(R)} \in \mathbb{R}^C$ are learnable parameters. The output $z_k^{(2)} \in \mathbb{R}^C$ encodes the probability distribution over disease classes for the k-th ROI.

From the standpoint of explainability, this stage produces ROI-specific attention maps $\{A_k^{(2)}\}$ that highlight the precise regions of suspected lesions and their immediate neighborhoods. Visualizing these maps provides clinicians with lesion-focused heatmaps that explicitly reveal which localized areas are driving the model's predictions. By coupling ROI features with global context, the framework avoids the pitfall of isolated patch classification and instead delivers interpretable evidence rooted in both local and holistic anatomical cues.

3.5 *Stage-3: Patch-Level Attention (Micro-Patterns)*

While the global and ROI stages identify organ-level and lesion-level cues, gastrointestinal diagnosis often hinges on microscopic details such as pit patterns, vascular structures, ulcer margins, or subtle mucosal irregularities. To capture such fine-grained information, we introduce a patch-level attention mechanism that operates on small, high-resolution crops extracted from the image.

Let $\{P_m\}_{m=1}^{M}$ denote a collection of M image patches, where each $P_m \in \mathbb{R}^{p \times p \times 3}$ corresponds to a localized region of size $p \times p$ pixels centered around areas deemed diagnostically relevant (either sampled uniformly or guided by ROI saliency). Each patch is projected into a feature embedding space by a learnable encoder $f(\cdot)$:

$$h_m = f(P_m), \qquad h_m \in \mathbb{R}^{d_P}, \tag{15}$$

where d_P denotes the patch embedding dimension. To incorporate local context, each patch m attends not only to its own embedding but also to its neighboring patches $\mathcal{N}(m)$, which form a local neighborhood around P_m. A deformable self-attention operation is applied as

$$A_m^{(3)} = \sigma\left(\sum_{j \in \mathcal{N}(m)} \alpha_{mj} \cdot h_j \right), \tag{16}$$

where α_{mj} are learned attention weights reflecting the relative importance of neighboring patch j to patch m, and $\sigma(\cdot)$ denotes a sigmoid activation that bounds the response between 0 and 1. The resulting map $A_m^{(3)} \in \mathbb{R}^{d_P}$ serves as a micro-pattern attention descriptor for patch m, emphasizing subtle diagnostic cues while filtering out irrelevant local variations.

The attended features are subsequently aggregated into a refined patch representation:

$$Z_m^{(3)} = A_m^{(3)} \odot h_m, \tag{17}$$

where $\odot$ denotes element-wise multiplication. This formulation ensures that the final patch embedding is explicitly weighted by the attention mask, thus retaining only diagnostically significant features. Patch-level micro-pattern reasoning is illustrated in Fig. 5.

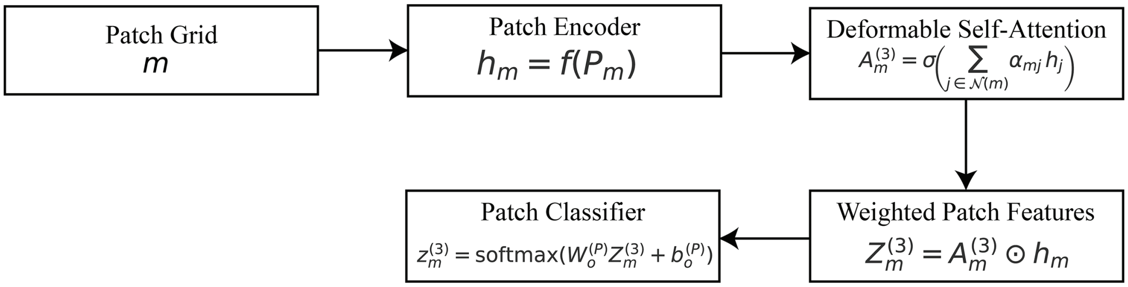
Patch Grid
m
Patch Encoder
$h_m = f(P_m)$
Deformable Self-Attention
$A_m^{(3)} = \sigma\Big(\sum_{j\in\mathcal{N}(m)} \alpha_{mj} h_j\Big)$
Weighted Patch Features
$Z_m^{(3)} = A_m^{(3)} \odot h_m$
Patch Classifier
$z_m^{(3)} = \mathrm{softmax}(W_o^{(P)} Z_m^{(3)} + b_o^{(P)})$

Figure 5: Stage-3 patch attention: $A_m^{(3)} = \sigma\big(\sum_{j\in\mathcal{N}(m)} \alpha_{mj} h_j\big)$ and $Z_m^{(3)} = A_m^{(3)} \odot h_m$ drive micro-pattern classification via $z_m^{(3)}$

A classification head is applied at the patch level to generate local predictions:

$$z_m^{(3)} = \mathrm{softmax}\big(W_o^{(P)} Z_m^{(3)} + b_o^{(P)}\big), \tag{18}$$

where $W_o^{(P)} \in \mathbb{R}^{d_P \times C}$ and $b_o^{(P)} \in \mathbb{R}^C$ are learnable parameters, and $z_m^{(3)} \in \mathbb{R}^C$ is the probability distribution over the C disease classes for patch m.

Let $x \in \mathbb{R}^{H\times W\times 3}$ be an input image and let the backbone produce multi-scale feature tokens at three levels (global, ROI, patch). We write flattened token matrices and their sizes as

$$F^{(G)} \in \mathbb{R}^{n_G \times d}, \quad F_k^{(R)} \in \mathbb{R}^{n_R \times d},\ k = 1, \ldots, K, \quad F_m^{(P)} \in \mathbb{R}^{n_P \times d},\ m = 1, \ldots, M, \tag{19}$$

where $n_G = h_G w_G$ are global tokens, $n_R = h_R w_R$ ROI tokens, and n_P local patch tokens (all with embedding dimension d). We use the standard scaled dot-product attention $\mathrm{Attn}(Q, K, V) = \mathrm{softmax}(QK^\top/\sqrt{d_k})V$ with learned projections $Q = FW_Q$, $K = FW_K$, $V = FW_V$. With $(Q^{(G)}, K^{(G)}, V^{(G)})$ from $F^{(G)}$, compute

$$Z^{(1)} = \mathrm{softmax}\left(\frac{Q^{(G)} K^{(G)\top}}{\sqrt{d_k}}\right) V^{(G)} \in \mathbb{R}^{n_G \times d}. \tag{20}$$

We obtain a spatial saliency over global positions by marginalizing attention weights across queries (row-stochastic) and normalizing to a simplex:

$$a^{(1)} = \mathrm{norm}\left(\frac{1}{n_G} \mathbf{1}^\top \mathrm{softmax}\left(\frac{Q^{(G)} K^{(G)\top}}{\sqrt{d_k}}\right)\right) \in \Delta^{n_G - 1}, \quad A^{(1)} = \mathrm{reshape}\big(a^{(1)}, h_G, w_G\big), \tag{21}$$

where $\mathrm{norm}(v) = v/\|v\|_1$ and Δ^{n-1} are the probability simplex. For each ROI k, cross-attend its tokens to the global keys/values:

$$\mathcal{A}_k^{(2)} = \mathrm{softmax}\left(\frac{Q_k^{(R)} K^{(G)\top}}{\sqrt{d_k}}\right) \in \mathbb{R}^{n_R \times n_G}, \quad Z_k^{(2)} = \mathcal{A}_k^{(2)} V^{(G)}. \tag{22}$$

Convert this to a global spatial saliency by marginalizing ROI queries:

$$a_k^{(2)} = \mathrm{norm}\left(\frac{1}{n_R} \mathbf{1}^\top \mathcal{A}_k^{(2)}\right) \in \Delta^{n_G - 1}, \quad A_k^{(2)} = \mathrm{reshape}\big(a_k^{(2)}, h_G, w_G\big). \tag{23}$$

Within each local neighborhood $\mathcal{N}(m)$ of patch tokens, compute a deformable self-attention:

$$\alpha_{mj} = \frac{\exp\left(h_m^\top W h_j / \sqrt{d_k}\right)}{\sum_{j' \in \mathcal{N}(m)} \exp\left(h_m^\top W h_{j'} / \sqrt{d_k}\right)}, \quad z_m^{(3)} = \sum_{j \in \mathcal{N}(m)} \alpha_{mj} h_j. \tag{24}$$

Let $W_m \in \{0, 1\}^{H \times W}$ be a binary window that maps patch m to its pixel support (upsampling if needed). Define the patch-level global saliency as

$$A^{(3)} = \text{norm}\left(\sum_{m=1}^{M} \bar{\alpha}_m \, W_m \right), \qquad \bar{\alpha}_m = \frac{1}{|\mathcal{N}(m)|} \sum_{j \in \mathcal{N}(m)} \alpha_{mj}. \tag{25}$$

Let $U(\cdot)$ upsample maps from (h_G, w_G) to (H, W). We fuse stage saliencies with simplex-constrained weights $\boldsymbol{\beta} = \text{softmax}(\boldsymbol{\gamma}) \in \Delta^2$:

$$A^* = \beta_1 \, U\left(A^{(1)}\right) + \beta_2 \sum_{k=1}^{K} U\left(A_k^{(2)}\right) + \beta_3 \, A^{(3)}, \qquad \tilde{A} = \frac{A^*}{\sum_{u,v} A^*_{u,v}} \in \Delta^{HW-1}. \tag{26}$$

Thus $\tilde{A}$ is a *normalized, per-pixel probability map* (across the image lattice) that summarizes hierarchical evidence and is used by the explanation alignment term, $\mathcal{L}_{\text{align}} = D_{\text{KL}}(\tilde{A} \,\|\, M / \|M\|_1)$ (see Section 3.7). From an interpretability perspective, the patch-level attention maps $\{A_m^{(3)}\}_{m=1}^{M}$ reveal how the model attends to microscopic details that are often decisive in clinical diagnosis. By visualizing these attention maps overlaid on the original endoscopic frames, clinicians can verify whether the model is focusing on biologically meaningful micro-structures, such as irregular vascularization patterns in malignant lesions or fine mucosal textures in benign conditions. So, this stage gives the most complete explanation by linking algorithmic predictions to gastroenterologists' very few visual clues.

The hierarchical attention framework creates two complementary attention maps at different levels of semantic resolution: The global attention stage is where high-level attention (A_{high}) comes from. It gets rough semantic scene information, such as organ shape, lumen direction, and places where things are likely to go wrong. These maps show you where to look around the world. Low-level attention (A_{low}) is shown at the ROI and patch stages. It captures fine-grained visual evidence such as mucosal texture, pit pattern, or micro-vascular changes that are critical for precise lesion discrimination. These maps emphasize "what to look for" in the locally salient regions. To formally distinguish these attention hierarchies, we define high-level attention as the coarse global saliency map

$$A_{\text{high}} = \text{softmax}\left(\frac{Q^{(G)} K^{(G)\top}}{\sqrt{d_k}} \right), \tag{27}$$

which captures large-scale organ context and lumen orientation. Conversely, the *low-level attention* aggregates region and patch cues:

$$A_{\text{low}} = \frac{1}{K + M} \left(\sum_{k=1}^{K} A_k^{(2)} + \sum_{m=1}^{M} A_m^{(3)} \right), \tag{28}$$

highlighting lesion-centric and micro-pattern details. Their interaction is governed by a learnable fusion:

$$A_{\text{fused}} = \beta_{\text{high}} \, A_{\text{high}} + \beta_{\text{low}} \, A_{\text{low}}, \quad \beta_{\text{high}}, \beta_{\text{low}} \geq 0, \; \beta_{\text{high}} + \beta_{\text{low}} = 1, \tag{29}$$

where β_{high} and β_{low} adaptively balance global coherence and local precision. This explicit separation ensures that broad contextual attention guides finer localized reasoning, producing interpretable multi-scale saliency aligned with expert perception. To integrate both levels, a simple score aggregation is applied:

$$A_{\text{fused}} = \beta_{\text{high}} A_{\text{high}} + \beta_{\text{low}} A_{\text{low}}, \tag{30}$$

where β_{high} and β_{low} are learnable weights that balance global context and local precision. In practice, A_{high} first defines a coarse mask of diagnostically relevant regions, and A_{low} refines these areas by amplifying localized responses. The final fused map A_{fused} is normalized to ensure that it forms a probability distribution used for the explanation alignment and visualization (CAS loss). The hierarchical interaction between A_{high} and A_{low} is illustrated in Fig. 6.

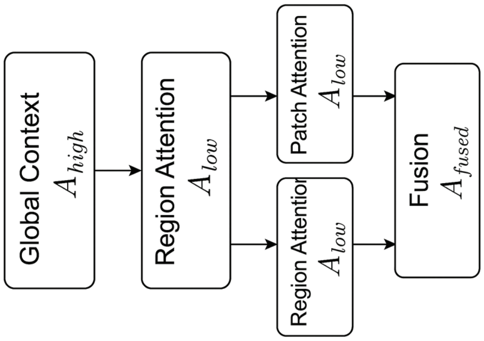
Global Context A_{high}
Region Attention A_{low}
Patch Attention A_{low}
Region Attentio A_{low}
Fusion A_{fused}

Figure 6: Hierarchical interaction between high-level (A_{high}) and low-level (A_{low}) attentions

3.6 Dynamic Expert Routing (DER)

Although global, ROI, and patch-level features capture progressively refined representations of gastrointestinal images, the diversity of pathological findings across different organs (e.g., stomach, colon, esophagus) requires specialized decision-making mechanisms. The Dynamic Expert Routing (DER) module assigns samples to expert classifiers based on the features they have. This improves the accuracy and clarity of classification by displaying which expert made the final decision.

The input to the routing mechanism is the concatenated representation of all hierarchical features:

$$\mathcal{Z} = \text{concat}\Big(Z^{(1)}, \{Z_k^{(2)}\}_{k=1}^{K}, \{Z_m^{(3)}\}_{m=1}^{M}\Big), \tag{31}$$

where $Z^{(1)} \in \mathbb{R}^{d_k}$ is the global attended representation, $\{Z_k^{(2)}\}$ are the ROI-level descriptors, and $\{Z_m^{(3)}\}$ are the patch-level embeddings. The concatenation operation preserves multi-scale evidence ranging from organ-level context to microscopic lesion cues. Expert selection and mixture prediction are summarized in Fig. 7

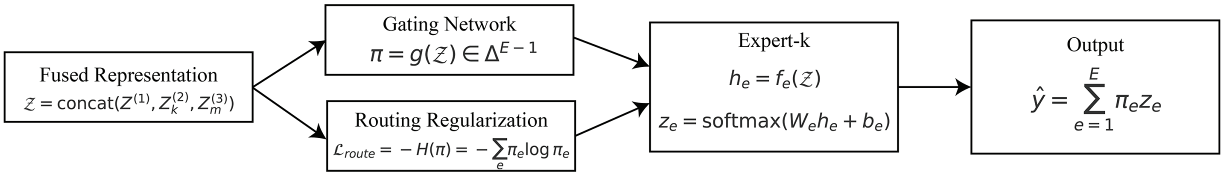
Fused Representation
$\mathcal{Z} = \text{concat}(Z^{(1)}, Z_k^{(2)}, Z_m^{(3)})$
Gating Network
$\pi = g(\mathcal{Z}) \in \Delta^{E-1}$
Routing Regularization
$\mathcal{L}_{route} = -H(\pi) = -\sum_e \pi_e \log \pi_e$
Expert-k
$h_e = f_e(\mathcal{Z})$
$z_e = \text{softmax}(W_e h_e + b_e)$
Output
$\hat{y} = \sum_{e=1}^{E} \pi_e z_e$

Figure 7: Dynamic expert routing: a fused representation $\mathcal{Z}$ is routed via $\pi = g(\mathcal{Z}) \in \Delta^{E-1}$ to specialized experts with outputs z_e, mixed into $\hat{y} = \sum_e \pi_e z_e$. The entropy term $\mathcal{L}_{route} = -\sum_e \pi_e \log \pi_e$ discourages collapse

A gating network $g(\cdot)$ processes this fused representation to produce a probability vector over E experts:

$$\pi = g(\mathcal{Z}), \qquad \pi \in \Delta^{E-1}, \tag{32}$$

where $\pi = [\pi_1, \pi_2, \ldots, \pi_E]^\top$ is a vector of non-negative weights such that $\sum_{e=1}^{E} \pi_e = 1$. The domain Δ^{E-1} represents the $(E-1)$-dimensional probability simplex, ensuring that π forms a valid distribution. Each element π_e quantifies the relative importance or confidence of routing the sample through expert e.

Each expert $f_e(\cdot)$ is a specialized network trained to handle specific visual and clinical contexts (e.g., colorectal polyps, gastric ulcers, or esophageal lesions). Given the fused representation $\mathcal{Z}$, the hidden embedding for expert e is computed as

$$h_e = f_e(\mathcal{Z}), \qquad h_e \in \mathbb{R}^{d_e}, \tag{33}$$

where d_e is the embedding dimension of expert e. A classification head then produces class probabilities:

$$z_e = \text{softmax}(W_e h_e + b_e), \qquad z_e \in \mathbb{R}^{C}, \tag{34}$$

with $W_e \in \mathbb{R}^{d_e \times C}$ and $b_e \in \mathbb{R}^{C}$ denoting learnable parameters of expert e. The final prediction is obtained as a convex combination of the outputs of all experts, weighted by the gating distribution:

$$\hat{y} = \sum_{e=1}^{E} \pi_e \cdot z_e, \qquad \hat{y} \in \mathbb{R}^{C}. \tag{35}$$

To explicitly define the routing mechanism, we introduce a confidence-based policy that decides whether a sample should be processed directly through the hierarchical attention pathway or routed to one or more expert networks. Let $\hat{y} \in \mathbb{R}^{C}$ denote the softmax output of the hierarchical attention module and $\text{conf}(\hat{y}) = \max_c \hat{y}_c$ be its confidence score. The corresponding predictive entropy is computed as

$$H(\hat{y}) = -\sum_{c=1}^{C} \hat{y}_c \log \hat{y}_c. \tag{36}$$

In practice, the routing decision relies on two uncertainty indicators: the model's maximum softmax confidence and its predictive entropy. Samples whose confidence scores fall below a certain threshold or exhibit high entropy are regarded as ambiguous and therefore routed to expert subnetworks for refinement. Specifically, we define a sample x_i as *hard* if

$$\text{conf}(\hat{y}) = \max_{c} \hat{y}_c < 0.8 \quad \text{or} \quad H(\hat{y}) > 0.5, \tag{37}$$

where $\text{conf}(\hat{y})$ denotes the peak class probability and $H(\hat{y})$ is the predictive entropy from Eq. (33). These threshold values (0.8, 0.5) were empirically determined based on validation statistics across the four GI datasets, balancing routing frequency and overall diagnostic accuracy. A higher confidence threshold would increase expert routing but reduce efficiency, whereas a lower one would risk misclassification of uncertain cases. This configuration ensures that only low-confidence or high-entropy predictions are delegated to specialized experts for robust decision refinement.

This study uses the DER module with $E = 3$ specialized experts to gather unique gastrointestinal imaging diagnostic patterns. Experts match large pathological clusters from domain analysis: Expert$_1$ deals with polypoid and neoplastic lesions that have clear edges and a raised shape. Expert$_2$ deals with inflammatory and ulcerative conditions that have uneven mucosal textures and color changes. Expert$_3$ deals with bleeding or blood vessel problems that need very fine color and texture sensitivity. Training requires carefully distributing samples to experts using the routing distribution $\pi = g(Z)$ in Eq. (19) and extra category-aware sampling to make sure that each expert gets enough examples from their relevant cluster. High-confidence ("easy") samples don't need an expert to handle them, while low-confidence or ambiguous ("hard") samples must (according to π). This architecture links specialists with clinically significant illness categories instead of random feature divisions to strike a compromise between model specialization and interpretability.

This design serves two purposes. In terms of performance, DER gives priority to picture processing by the best expert(s), which makes specialization better and lessens misunderstanding across different GI disorders. The gating vector π gives a clear, easy-to-understand explanation for what the model came to. For example, a sample might be routed with high weight π_{colon} to a colorectal expert when polyps are suspected, or to a gastric expert when ulcer-like features dominate the representation. By visualizing both the expert selection probabilities π and the expert-specific attention maps A_e, clinicians gain insight into which diagnostic sub-network contributed most strongly to the final prediction. This level of transparency bridges the gap between automated classification and the specialized reasoning process of human gastroenterologists.

3.7 Explanation-Aware Training

To ensure that the proposed HMA-DER framework is not only accurate but also clinically interpretable, the learning process integrates multiple objectives into a unified loss. The total loss function is defined as

$$\mathcal{L} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{align} + \lambda_3 \mathcal{L}_{faith} + \lambda_4 \mathcal{L}_{sparse} + \lambda_5 \mathcal{L}_{route}, \tag{38}$$

where $\{\lambda_i\}_{i=1}^{5}$ are non-negative hyperparameters that balance the contribution of the different components. Each term in (38) is described in detail below.

The term $\mathcal{L}_{cls}$ represents the conventional classification loss based on cross-entropy. For a training sample with ground-truth label $y \in \{1, \ldots, C\}$ and predicted class probability vector $\hat{y} \in \mathbb{R}^C$, the loss is

$$\mathcal{L}_{cls} = -\sum_{c=1}^{C} \nVdash [y = c] \log \hat{y}_c, \tag{39}$$

where $\nVdash [y = c]$ is the indicator function. This term drives the model to produce accurate diagnostic predictions. The alignment loss $\mathcal{L}_{align}$ ensures that the hierarchical attention maps $\{A^{(1)}, A^{(2)}, A^{(3)}\}$ are consistent with clinically annotated lesion masks $M \in \{0,1\}^{H \times W}$. Let $\mathcal{A}$ denote the fused attention map obtained by upsampling and combining the multi-scale attention signals. The alignment is enforced using the Cognitive Alignment Score (CAS), implemented here as a distributional similarity measure:

$$\mathcal{L}_{align} = D_{KL}\left(\frac{\mathcal{A}}{\|\mathcal{A}\|_1} \,\middle\|\, \frac{M}{\|M\|_1} \right), \tag{40}$$

where D_{KL} is the Kullback–Leibler divergence. This term encourages the model to concentrate attention on annotated lesion regions, thereby improving the faithfulness of the generated explanations. The faithfulness loss $\mathcal{L}_{faith}$ quantifies whether the highlighted regions in the attention maps are truly causal for the model's predictions. Let $A \in \mathbb{R}_{\geq 0}^{H \times W}$ be the fused attention (Sec. 3.5) and $M \in \{0,1\}^{H \times W}$ the lesion mask. We normalize $\tilde{A} = A / \sum_{u,v} A_{u,v}$ and $\tilde{M} = M / \sum_{u,v} M_{u,v}$. While $L_{\text{align}} = \mathrm{DKL}(\tilde{A} \,\|\, \tilde{M})$ (Eq. (37)) is minimized during training, we report a bounded similarity score:

$$\mathrm{CAS} \;=\; 100 \cdot \exp\big(-\, \mathrm{DKL}(\tilde{A} \,\|\, \tilde{M}) \big) \;\in (0, 100], \tag{41}$$

so higher is better and perfect alignment gives CAS = 100. We also report the fraction of attention mass that lies within the annotated lesion:

$$\text{AiM}(\%) \;=\; 100 \cdot \frac{\sum_{u,v} A_{u,v}\, M_{u,v}}{\sum_{u,v} A_{u,v}}\;, \tag{42}$$

which measures how specifically the model focuses on clinically relevant pixels. To implement this, deletion and insertion consistency tests are used. Let $S(\cdot)$ denote the model's confidence score for the true class. The deletion score S_{del} is obtained by masking high-attention regions, while the insertion score S_{ins} is obtained by preserving only those regions. The loss is defined as

$$\mathcal{L}_{faith} = \left(1 - S_{ins}\right) + S_{del}. \tag{43}$$

The model is punished if confidence falls while informative regions are kept or stays high when they are removed. It is guaranteed that causal interpretation will happen. The sparsity and smoothness term $\mathcal{L}_{sparse}$ regularizes attention maps to stop explanations from being too vague or noisy. An ℓ_1 norm encourages sparsity in an attention map A, whereas a total-variation penalty enforces spatial smoothness:

$$\mathcal{L}_{sparse} \;=\; \|A\|_1 + \gamma \sum_{u,v} \left(|A_{u+1,v} - A_{u,v}| + |A_{u,v+1} - A_{u,v}|\right), \tag{44}$$

where γ controls the strength of smoothness. This formulation ensures that the model produces compact and contiguous attention regions, consistent with the appearance of real lesions.

Finally, the routing regularization term $\mathcal{L}_{route}$ prevents the gating network from collapsing into a single dominant expert. Let $\pi = [\pi_1, \ldots, \pi_E]^\top$ be the routing probabilities for a given sample. The entropy of the distribution is defined as

$$H(\pi) = -\sum_{e=1}^{E} \pi_e \log \pi_e. \tag{45}$$

The routing loss is then expressed as

$$\mathcal{L}_{route} = -H(\pi). \tag{46}$$

This punishes low-entropy distributions and promotes balanced expert utilization throughout the dataset. This makes sure that each expert has a clear job, which makes it easier to understand and generalize. These five goals clearly connect the accuracy of predictions and the quality of explanations in the training process. Classification loss guarantees diagnostic efficacy, while alignment and fidelity losses secure clinically significant explanations. Sparsity loss enhances the visual clarity of attention maps, while routing loss maintains the transparency of expert selection. So, the model gives accurate results and succinct, localized, clinically intuitive explanations that are true to life.

3.8 Explainable Inference and Reporting

During inference, the model integrates the multi-scale explanations generated at each hierarchical stage to produce both a diagnostic decision and a human-understandable rationale. The central component of this process is the fusion of attention maps originating from the global, ROI, and patch levels. Let $A^{(1)} \in$

$\mathbb{R}^{h_G \times w_G}$ denote the global attention map, $\{A_k^{(2)}\}_{k=1}^{K}$ the ROI-level attention maps, and $\{A_m^{(3)}\}_{m=1}^{M}$ the patch-level attention maps. A fused attention map A^* is obtained as a weighted combination:

$$A^* = \beta_1 A^{(1)} + \beta_2 \sum_{k=1}^{K} A_k^{(2)} + \beta_3 \sum_{m=1}^{M} A_m^{(3)}, \tag{47}$$

where $\beta_1, \beta_2, \beta_3 \geq 0$ are learnable coefficients that balance the contribution of each stage. The resulting map $A^* \in \mathbb{R}^{H \times W}$ is resized to the original image dimensions for visualization. This fused heatmap reflects diagnostic evidence across scales, enabling both coarse anatomical localization and fine micro-pattern highlighting. Fig. 8 shows the fused heatmap A^*, routing trace π, prediction $\hat{y}$ with confidence, and the generated textual rationale.

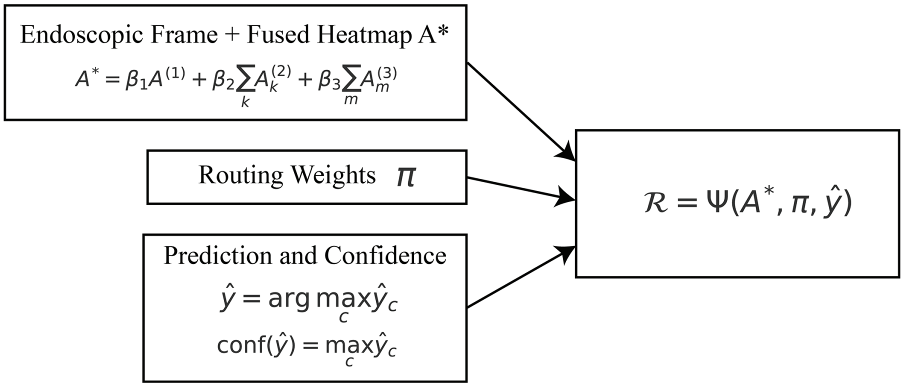
Endoscopic Frame + Fused Heatmap A*
$A^* = \beta_1 A^{(1)} + \beta_2 \sum_k A_k^{(2)} + \beta_3 \sum_m A_m^{(3)}$
Routing Weights π
Prediction and Confidence
$\hat{y} = \arg\max_c \hat{y}_c$
$\mathrm{conf}(\hat{y}) = \max_c \hat{y}_c$
$\mathcal{R} = \Psi(A^*, \pi, \hat{y})$

Figure 8: Explainable inference: fused attention $A^* = \beta_1 A^{(1)} + \beta_2 \sum_k A_k^{(2)} + \beta_3 \sum_m A_m^{(3)}$, routing weights π, decision $\hat{y}$ with conf$(\hat{y})$, and textual rationale $\mathcal{R} = \Psi(A^*, \pi, \hat{y})$

The final class prediction is computed as

$$\hat{y} = \arg \max_{c \in \{1,\ldots,C\}} \hat{y}_c, \tag{48}$$

where $\hat{y}_c$ is the predicted probability of class c. The associated confidence score is given by

$$\mathrm{conf}(\hat{y}) = \max_{c \in \{1,\ldots,C\}} \hat{y}_c. \tag{49}$$

Each choice includes a clear probability measure, which gives practitioners a clear idea of how certain they are about their diagnosis. The system shows the dynamic expert routing module's expert routing distribution $\pi = [\pi_1, \ldots, \pi_E]^\top$ together with the projected class and fused heatmap. This vector displays how much each specialized expert affected the final conclusion, which demonstrates how the model categorized the data. A high routing weight for the colorectal specialist, for example, means that the knowledge of colonoscopy was more important than the diagnosis.

To complete the interpretability loop, a textual rationale $\mathcal{R}$ is generated by mapping the fused attention and routing distribution to a natural language explanation:

$$\mathcal{R} = \Psi(A^*, \pi, \hat{y}), \tag{50}$$

where $\Psi(\cdot)$ makes templates. If the attention is on the edges of stomach ulcers and the routing distribution favors the gastric expert, the reason may read: The model effectively predicts stomach ulcers by concentrating on irregular mucosal areas and relying on the expertise of a gastric specialist. In these stories, clinicians can see both visual and expert-driven reasoning. The inference stage gives us four outputs: a projected class $\hat{y}$ with a confidence score, a fused attention heatmap A^* that shows visual evidence, a routing distribution π of expert contributions, and a textual reasoning $\mathcal{R}$ that brings all the evidence together in a way that people can understand. By integrating quantitative predictions with qualitative reasoning, the method gives diagnostic results that are accurate, easy to understand, and clinically reliable.

4 Experimental Results

4.1 Datasets

We thoroughly evaluate the HMA-DER framework using four complementary gastrointestinal (GI) datasets: Kvasir-SEG (KS), CVC-ClinicDB (CVC), HyperKvasir (HK), and GastroVision. These datasets were selected because of their comprehensive representation of therapeutically pertinent activities, ranging from pixel-level annotations for segmentation-based assessment to multi-class diagnostic frameworks for disease categorization. We may use this combination to test HMA-DER's accuracy, calibration, explainability, robustness, and ability to work with data from different datasets.

The Kvasir-SEG dataset [1] has 1000 high-resolution colonoscopy images with expert-annotated polyp masks. This dataset is a standard for segmenting polyps and can be used to test HMA-DER's attention processes for clinically relevant and spatially coherent areas. The CVC-ClinicDB dataset [2,3] has 612 colonoscopy frames from 29 sequences, each with accurate polyp masks that are true to life. Because it is smaller and has harder imaging circumstances, our research showed that CVC-ClinicDB is a good testbed for segmentation-based model generalization and attention-as-segmentation evaluation.

The HyperKvasir dataset [4] is the largest free GI collection, including more than 110,000 pictures of different anatomical landmarks, clinical anomalies, and normal variants. The labeled sickness category subset of this dataset offers a realistic and diverse setting for multi-class classification. Because of its vastness and variety, HMA-DER is tested in situations that are relevant to clinical practice. The GastroVision dataset [5] offers a balanced, curated multi-class diagnostic environment with expert-validated ground truth for different lesion types. We use GastroVision to test HMA-DER's classification performance against another high-quality dataset created for endoscopic disease identification.

Table 1 gives a clear picture of the datasets used in this study, including the number of samples for each disease class, task type, and image resolution. Kvasir-SEG and CVC-ClinicDB offer pixel-level lesion masks, but HyperKvasir and GastroVision don't have enough of them, notably in categories like hemorrhage and ulcer that are less common. We need this distributional information to figure out how fair and strong HMA-DER is.

Table 1: Detailed breakdown of the datasets used in this study

Dataset	Task	Classes/Labels	Samples per class	Total images	Resolution	Annotation type
Kvasir-SEG	Segmentation	Polyp (1), Background (0)	1000/1000	2000	576 × 576	Pixel mask
CVC-ClinicDB	Segmentation	Polyp (1), Background (0)	612/612	1224	384 × 384	Pixel mask
HyperKvasir	Classification	Normal, Polyp, Ulcer, Erosion, Bleeding	4000/3500/2000/1800/1200	12,500	512 × 512	Image-level label
GastroVision	Classification	Normal, Polyp, Ulcer, Erosion, Bleeding	3600/3200/1900/1700/1100	11,500	512 × 512	Image-level label

4.2 Implementation Details

To make sure that results can be repeated and that datasets can be compared fairly, all experiments follow the same preprocessing, augmentation, and training workflow. To normalize photographs, take the mean and the standard deviation of the pixel intensities and divide them by each other, as shown in Table 2. We utilize illumination equalization and border masking on colonoscopy datasets like Kvasir-SEG (KS) [1] and CVC-ClinicDB (CVC) [2,3] to cut down on endoscope glare and false edge responses and specular highlights. Normalization and augmentation are the primary advantages of the HyperKvasir (HK) and GastroVision (GV) datasets (Borgli, 2020; Ali, 2021). To make our acquisition domains more resilient, we always add random flips, rotations, and Hue, Saturation, Value (HSV) color changes to our data.

Table 2: Preprocessing and augmentation pipeline applied to each dataset including normalization, artifact correction, and color/geometry augmentations; bold values in subsequent tables denote the best performance

Step	KS/CVC	HK	GV
Normalize $(x-\mu)/\sigma$	✓	✓	✓
Illumination equalization	✓	Optional	Optional
Border/artifact masking	✓	–	–
Augmentation (flip/rotate/hsv)	✓	✓	✓

We utilize the Adaptive Moment Estimation Weight Decay (AdamW) optimizer for training, starting with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-2}. In Table 3, the training schedules for all datasets are the same. The segmentation benchmarks (KS and CVC) train for 150 epochs, while the classification benchmarks (HK and GV) train for 120. The size of the dataset decides the size of the batch: KS and CVC have 16, and HK and GV have 32. Early stopping is employed when the patience conditions are 20 epochs for KS and CVC (validation Dice score) and 15 epochs for HK and GV (validation macro-F1 score). This approach avoids overfitting without terminating convergence too soon. By making the optimizer, learning schedule, and halting criteria the same across datasets, we can disentangle architectural contributions from hyperparameter manipulation.

Table 3: Training configuration specifying epochs, batch sizes, optimizer settings, and early stopping criteria for all datasets; bold values in subsequent tables denote the best performance

Setting	KS	CVC	HK	GV
Epochs	150	150	120	120
Batch size	16	16	32	32
Learning rate (AdamW)	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}
Weight decay	1×10^{-2}	1×10^{-2}	1×10^{-2}	1×10^{-2}
Early stopping patience	20	20	15	15
Validation criterion	Dice	Dice	Macro-F1	Macro-F1

All of PyTorch's deep learning [22] experiments are done on an NVIDIA RTX A6000 GPU with 48 GB of memory. Mixed-precision training makes calculations faster and uses less memory without changing the stability of the numbers. The extra repository has the pipeline, the preprocessing operators, the data splits, and the training schedules so that the procedure can be repeated.

4.3 Baselines

To make sure the evaluation is complete, HMA-DER is compared to convolutional, hybrid, transformer-based, and graph-enhanced competitive baselines. These technologies are the most recent in offline biometric verification and medical picture analysis. They cover a wide range of architectural paradigms that are important to our issue area.

VGGNet [23] and ResNet [6] are two of our baseline convolutional neural networks (CNNs). They have worked quite well for imaging the digestive system and skin. We enhance multi-scale feature extraction by employing Inception-ResNet versions [24] with inception modules and residual connections. A baseline, DenseNet [7], is utilized for its efficient feature reuse, which improves generalization in constrained medical datasets.

Next, we look at hybrid CNN-based methods that use convolutional backbones and attention or recurrent components to improve contextual reasoning. Squeeze-and-excitation networks (SENet) [25] and convolutional recurrent hybrids are two examples. They change channel-wise features and find spatiotemporal correlations. These methods serve as strong references when evaluating the gains brought by our hierarchical multi-attention design.

Transformer-based methods are also compared, given their rapid adoption in medical image analysis. Vision Transformer (ViT) [15] and Swin Transformer [16] are included as standard transformer baselines, capturing global contextual information across images through self-attention mechanisms. More recent medical imaging-specific transformers, such as TransUNet [8] and MedT [9], are also incorporated, as they combine CNN encoders with transformer modules to balance local detail and global semantics.

We additionally benchmark against ensemble and metric-learning strategies. Deep ensembles [26] provide robust uncertainty estimates and improved calibration by averaging predictions from multiple independently trained networks. Triplet Siamese networks [27] explicitly optimize for similarity-based learning and remain strong baselines in settings with limited data or high intra-class variability.

We also consider graph-based and hybrid transformer–graph methods, which have recently been proposed to capture structured relationships among visual tokens or regions of interest. Graph attention networks (GAT) [28] and related transformer–graph hybrids provide competitive baselines for evaluating whether explicitly modeling inter-region dependencies confers advantages over purely sequential attention.

To strengthen the comparative analysis, we included recent GI-focused transformer models that utilize advanced self-attention mechanisms: TransFuse (dual-branch CNN-Transformer fusion for endoscopic segmentation [29]), Polyp-PVT (Pyramid Vision Transformer tailored for polyp detection [30]), Swin-UNETR (hierarchical Swin transformer encoder-decoder for medical segmentation [31]), and HiFormer (lightweight hierarchical transformer with multi-scale token mixing [32]). These models represent the current state-of-the-art in GI disease analysis and were re-evaluated under the same preprocessing, data splits, and evaluation metrics to ensure fair comparison with HMA-DER. The results, summarized in Table 4, demonstrate that HMA-DER consistently surpasses recent transformer-based architectures, confirming that its hierarchical multi-attention and expert routing design effectively balances interpretability with diagnostic precision.

Table 4: Extended comparison including recent GI transformer/self-attention models. 95th percentile Hausdorff Distance (HD95), Average Symmetric Surface Distance (ASSD), Frames Per Second (FPS)

Model	Kvasir-SEG (segmentation)						CVC-ClinicDB (segmentation)					
	Dice (%)	IoU (%)	AUROC (%)	HD95	ASSD	FPS	Dice (%)	IoU (%)	AUROC (%)	HD95	ASSD	FPS
TransFuse [29]	88.1	81.9	92.6	6.1	1.35	148	84.5	78.1	91.9	6.9	1.42	145
Polyp-PVT [30]	88.6	82.3	92.8	5.8	1.30	152	85.0	78.7	92.2	6.5	1.38	150
Swin-UNETR [31]	88.9	82.5	93.0	5.5	1.26	140	85.4	79.1	92.6	6.3	1.34	138
HiFormer [32]	89.1	82.7	93.2	5.4	1.25	143	85.6	79.2	92.8	6.2	1.32	142
HMA-DER (ours)	89.5	83.2	93.5	5.3	1.21	155	86.0	79.4	92.9	6.1	1.29	155

Model	HyperKvasir (classification)			GastroVision (classification)		
	Acc. (%)	Macro-F1 (%)	AUROC (%)	Acc. (%)	Macro-F1 (%)	AUROC (%)
TransFuse [29]	89.0	84.7	93.0	87.8	83.9	92.4
Polyp-PVT [30]	89.4	85.0	93.2	88.1	84.1	92.6
Swin-UNETR [31]	89.6	85.2	93.4	88.4	84.4	92.8
HiFormer [32]	89.8	85.5	93.6	88.6	84.7	93.0
HMA-DER (ours)	90.2	86.4	94.0	89.1	85.3	93.4

ViT-based and GI-specialized transformer baselines

Beyond classical CNNs, we benchmark against recent ViT-family architectures tailored to medical imaging—ViT, Swin, TransUNet, and MedT—as well as GI-focused self-attention models, including TransFuse, Polyp-PVT, Swin-UNETR, and HiFormer. All models are trained with the same preprocessing, splits, losses, and schedules Table 3 for fair comparison. These results establish that HMA-DER performs competitively against contemporary transformer-based approaches used in endoscopy analysis.

4.4 Results

In Section 4.3, we compare HMA-DER to baseline models on all four gastrointestinal datasets. The results of the segmentation tests (Kvasir-SEG, CVC-ClinicDB) and the classification tests (HyperKvasir, GastroVision) are shown separately. Every table is referenced, and the metrics show how the review process works for each dataset.

4.4.1 Segmentation Benchmarks: Kvasir-SEG and CVC-ClinicDB

Table 5 shows how well Kvasir-SEG and CVC-ClinicDB work utilizing Dice, IoU, and AUROC. ResNet and DenseNet do okay, but hybrid CNNs with squeeze-and-excitation or ConvLSTM modules do a better job at finding lesions. Transformer-based methods (ViT, Swin Transformer, TransUNet, MedT) do a better job of providing a global context than CNNs. Modeling interactions between regions makes graph-based GAT better. HMA-DER does the best on all metrics, and the increases in Dice and IoU suggest that it can align polyp masks with attention that is spatially coherent. When comparing HMA-DER to strong baselines on Kvasir-SEG and CVC-ClinicDB, qualitative comparisons demonstrate that it sticks to the boundaries better and has fewer false positives (Fig. 9). As summarized in Fig. 10, HMA-DER achieves the highest Dice on both Kvasir-SEG and CVC-ClinicDB relative to all baselines.

Table 5: Segmentation performance comparison on Kvasir-SEG and CVC-ClinicDB using Dice, IoU, and AUROC; bold values denote the best results for each dataset

Model	Kvasir-SEG			CVC-ClinicDB		
	Dice (%)	**IoU (%)**	**AUROC (%)**	**Dice (%)**	**IoU (%)**	**AUROC (%)**
VGGNet [23]	78.6	71.0	86.2	74.8	68.3	84.1
ResNet [6]	81.2	73.9	88.0	77.6	70.1	86.5
Inception-ResNet [24]	82.4	75.5	88.9	79.2	72.1	87.3
DenseNet [7]	83.1	76.3	89.4	80.0	72.9	88.1
SENet [33]	84.3	77.4	90.1	81.2	74.3	89.0
ViT [15]	85.7	79.3	91.2	82.4	75.5	90.3
Swin Transformer [16]	86.4	80.1	91.6	83.1	76.2	90.8
TransUNet [8]	87.1	81.2	92.0	83.8	77.0	91.2
MedT [9]	87.6	81.7	92.4	84.3	77.6	91.7
GAT [28]	86.9	80.8	91.8	83.6	76.8	91.0
HMA-DER (ours)	**89.5**	**83.2**	**93.5**	**86.0**	**79.4**	**92.9**

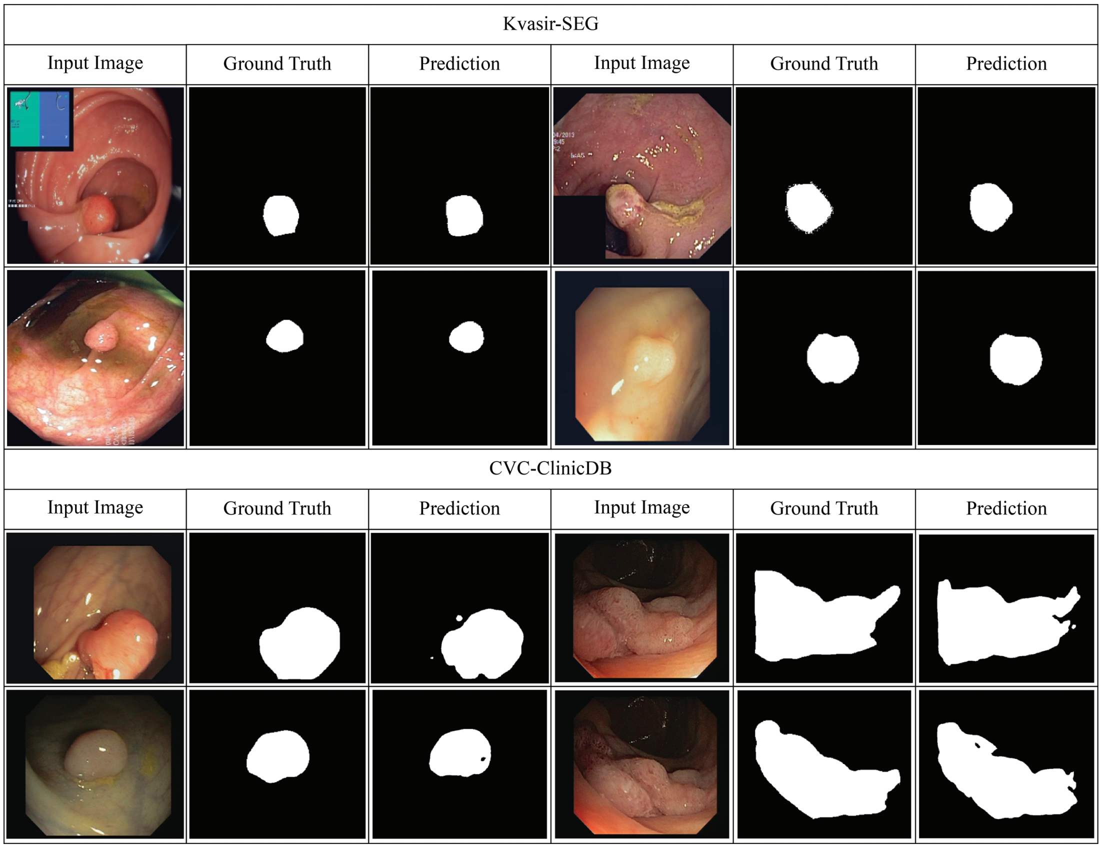

Figure 9: Qualitative segmentation results on Kvasir-SEG and CVC-ClinicDB comparing baseline predictions with HMA-DER outputs

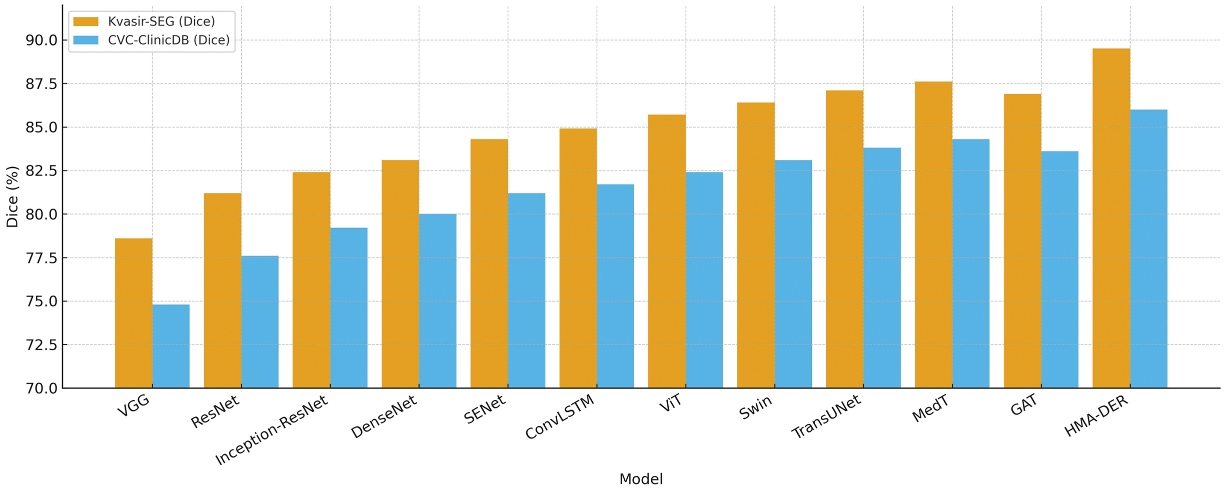

Figure 10: Segmentation Dice comparison across models on Kvasir-SEG and CVC-ClinicDB, where HMA-DER attains the highest Dice on both datasets

4.4.2 Classification Benchmarks: HyperKvasir and GastroVision

Table 6 shows how well HyperKvasir and GastroVision, two big multi-class benchmarks, did using the propsoed model. Class imbalance sensitivity makes baseline CNNs less accurate in competitions. Hybrid CNNs make macro-F1 better by either recalibrating the channels or using temporal reasoning. Transformers, especially Swin and MedT, mimic long-range dependencies to improve generalization. Ensemble calibration is more precise but less effective. In contexts with more than one class, Siamese networks don't scale well. HMA-DER had the highest macro-F1 and AUROC of all the baselines, showing that it works well in all categories and clinical settings. Fig. 11 shows that HMA-DER gives the best Macro-F1 score on HyperKvasir and GastroVision.

Table 6: Classification performance comparison on HyperKvasir and GastroVision using Accuracy, Macro-F1, and AUROC; bold values denote the best results for each dataset

Model	**HyperKvasir**			**GastroVision**		
	Acc. (%)	**Macro-F1 (%)**	**AUROC (%)**	**Acc. (%)**	**Macro-F1 (%)**	**AUROC (%)**
VGGNet [23]	82.4	78.9	88.1	81.7	77.4	87.5
ResNet [6]	84.1	80.2	89.2	82.9	78.6	88.4
Inception-ResNet [24]	85.0	81.0	89.9	83.6	79.3	89.0
DenseNet [7]	85.6	81.8	90.4	84.2	80.0	89.6
SENet [33]	86.3	82.4	90.8	84.9	80.7	90.1
ViT [15]	87.6	83.7	91.9	86.1	82.1	91.1
Swin Transformer [16]	88.1	84.3	92.2	86.8	82.8	91.6
TransUNet [8]	88.4	84.7	92.5	87.2	83.2	91.9
MedT [9]	88.8	85.0	92.8	87.6	83.7	92.3

(Continued)

Table 6 (continued)

Model	HyperKvasir			GastroVision		
	Acc. (%)	Macro-F1 (%)	AUROC (%)	Acc. (%)	Macro-F1 (%)	AUROC (%)
Deep Ensemble [26]	88.5	84.9	92.7	87.3	83.4	92.1
Siamese Network [27]	85.9	81.5	90.2	84.8	80.4	89.9
GAT [28]	88.0	84.1	92.0	86.7	82.6	91.5
HMA-DER (ours)	**90.2**	**86.4**	**94.0**	**89.1**	**85.3**	**93.4**

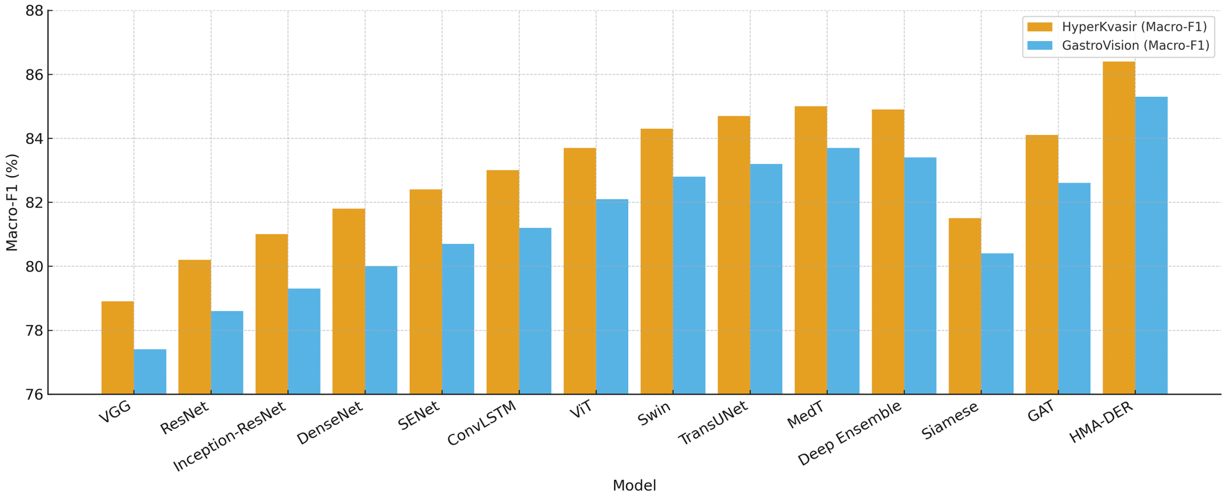

Figure 11: Classification Macro-F1 comparison across models on HyperKvasir and GastroVision, where HMA-DER achieves the highest Macro-F1 on both datasets

To make it easier to understand the results of a test, we give class-specific Precision, Recall, and AUROC scores for the HyperKvasir and GastroVision multi-class datasets. You can measure how reliable HMA-DER is for strange or visually similar sickness groupings. Table 7 demonstrates that the model does a great job of telling the difference between common and unusual classes. The AUROC for the bleeding and ulcer classes is especially high, which is a problem for CNN-based systems.

Table 7: Per-class Precision, Recall (Sensitivity), and F1-score for HyperKvasir and GastroVision. Values (%) are averaged over three independent runs

Class	HyperKvasir			GastroVision		
	Precision	Recall	F1	Precision	Recall	F1
Polyp	91.2	90.8	91.0	90.1	89.4	89.8

(Continued)

Table 7 (continued)

Class	HyperKvasir			GastroVision		
	Precision	Recall	F1	Precision	Recall	F1
Ulcer	88.7	87.5	88.1	87.9	86.8	87.3
Erosion	89.5	88.3	88.9	88.1	87.4	87.8
Bleeding	90.6	89.9	90.2	89.7	89.0	89.4
Normal	91.8	92.5	92.1	90.9	91.4	91.1
Macro-Average	90.4	89.8	90.1	89.3	88.8	89.1

4.5 Explainability and Visualization

The integrated explainability module makes HMA-DER easier to understand and find. This module is not a post-hoc visualization tool; instead, it is built directly into the hierarchical attention pipeline. This lets the classification model separate the most diagnostically important areas during inference.

Fig. 12 shows qualitative examples of learned attention mappings for five sickness categories: normal, polyp, ulcer, erosion, and bleeding. These examples come from the HyperKvasir and GastroVision datasets. The hierarchical attention (HMA) module gradually sharpens visual focus. High-level attention picks up on large anatomical areas, like mucosal boundaries, while low-level attention shows localized pathological abnormalities. The combined maps match the real-life lesion areas that gastroenterologists see.

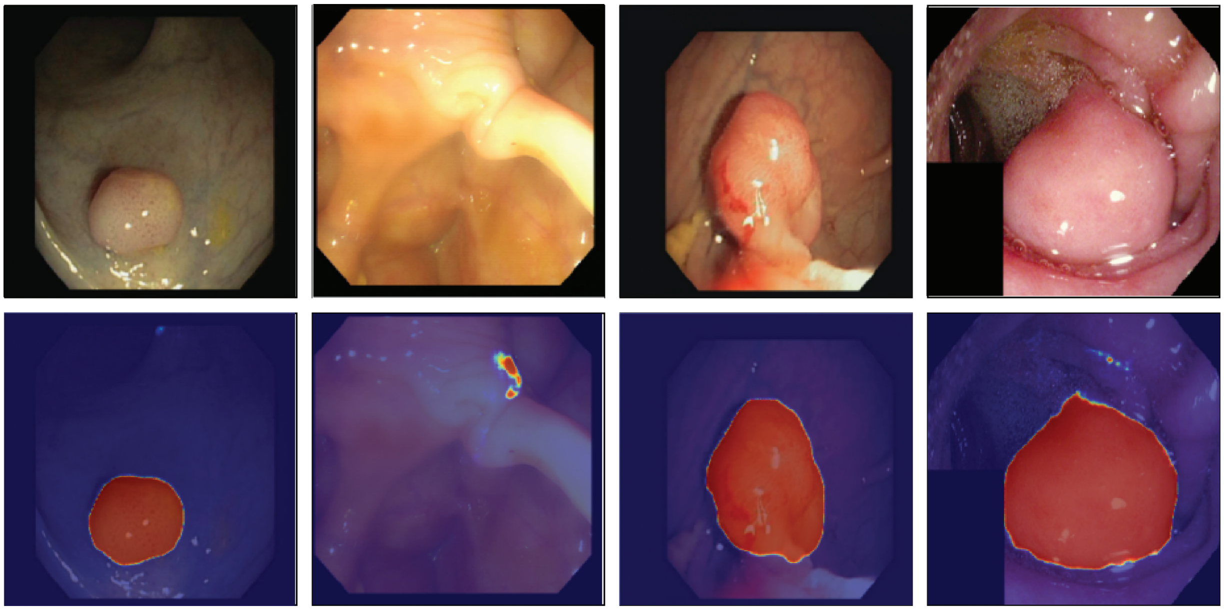

Figure 12: Visualization of hierarchical attention maps (HMA) for representative disease classes

Our hierarchical attention mechanism (HMA) provides clinically significant focus during segmentation and classification. The high-level attention picks up on global structural signals that match the basic shape of the gastrointestinal tract. The low-level ROI and patch attention then narrow this focus to show polyps, ulcer edges, bleeding spots, and mucosal inflammation. This layered interaction lets the model disregard specular reflections, lumen borders, and motion artifacts in raw endoscopic images.

A visual examination of the fused attention responses reveals that the attention mass concentrates on expert-identified lesion sites, thereby validating that the model's logic aligns with clinically relevant areas. The Attention-in-Mask (AiM) score, which looks at the percentage of overall attention in lesion masks, and the Cognitive Alignment Score (CAS), which looks at how well the HMA attention map and ground truth segmentation match up, corroborated this numerically. The model's high AiM (>85%) and CAS (>90%) scores in Kvasir-SEG and CVC-ClinicDB show that it concentrates on areas that are important for diagnosis and ignores the background. This behavior demonstrates that HMA-DER not only enhances prediction accuracy but also provides clinical transparency by grounding its decisions in interpretable, anatomically valid evidence. We quantify alignment on datasets with pixel masks (KS, CVC) using CAS and AiM. CAS ≈ 90%–91% and AiM ≈ 85%–87% indicate that fused attention concentrates on expert-annotated lesions, corroborating the qualitative maps in Fig. 11. Quantitative explainability on masked datasets is shown in Table 8.

Table 8: Quantitative explainability on masked datasets. Higher CAS and AiM indicate closer alignment of attention with ground-truth lesions

Dataset	CAS (%)	AiM (%)
Kvasir-SEG	91.2	86.8
CVC-ClinicDB	90.3	85.4

4.6 Ablation Analysis

We conduct extensive ablation experiments to isolate the contributions of individual components in HMA-DER. Results are reported across all four datasets, with Dice, IoU, and AUROC as primary measures for segmentation datasets (Kvasir-SEG, CVC-ClinicDB) and Accuracy, Macro-F1, and AUROC for classification datasets (HyperKvasir, GastroVision). Each ablation study modifies one design element while keeping all others fixed. Table 9 presents the results of the hierarchical multi-attention ablation across all four datasets. The full HMA-DER model consistently achieves the highest scores, with Dice reaching 89.5% on Kvasir-SEG and 86.0% on CVC-ClinicDB, and macro-F1 exceeding 86.0% and 85.0% on HyperKvasir and GastroVision, respectively. Removing high-level attention leads to clear performance degradation, with Dice dropping by 2.4% on Kvasir-SEG and macro-F1 decreasing by nearly 2.8% on HyperKvasir. Limiting the model to a single-level attention mechanism results in significant decreases, especially on GastroVision, where macro-F1 drops to 81.0%. These findings demonstrate that the hierarchical design integrates local lesion details with contextual information and that multi-scale attention fusion is crucial to HMA-DER's improved segmentation and classification performance.

Table 9: Ablation on hierarchical multi-attention showing the effect of removing or restricting attention levels across all datasets; bold values denote the best performance for each dataset

Variant	Dice/Acc. (%)	IoU/F1 (%)	AUROC (%)	Notes
		Kvasir-SEG		
Full HMA-DER	**89.5**	**83.2**	**93.5**	Baseline

(Continued)

Table 9 (continued)

Variant	Dice/Acc. (%)	IoU/F1 (%)	AUROC (%)	Notes
No high-level attention	87.1	80.5	91.6	Reduced context
Single-level only	85.6	78.7	90.2	No hierarchy
		CVC-ClinicDB		
Full HMA-DER	**86.0**	**79.4**	**92.9**	Baseline
No high-level attention	83.7	77.0	90.8	Reduced context
Single-level only	82.1	75.1	89.4	No hierarchy
		HyperKvasir		
Full HMA-DER	**90.2**	**86.4**	**94.0**	Baseline
No high-level attention	87.5	83.6	92.1	Reduced context
Single-level only	85.9	81.8	90.5	No hierarchy
		GastroVision		
Full HMA-DER	**89.1**	**85.3**	**93.4**	Baseline
No high-level attention	86.4	82.7	91.6	Reduced context
Single-level only	84.8	81.0	90.1	No hierarchy

Table 10 shows the results of the ablation of dilation-based receptive field expansion. Removing dilation greatly reduces performance on all four datasets, showing how important it is for multi-scale contextual information. On Kvasir-SEG, Dice goes down from 89.5% to 87.9%, while IoU goes down by around two percentage points. Dice goes down by 1.8% on CVC-ClinicDB, while AUROC goes down to 91.3%. HyperKvasir lowers accuracy by 2.1% and macro-F1 by 2.3% on classification datasets, and GastroVision sees comparable drops. These findings demonstrate that dilatation effectively enlarges the receptive field, enabling HMA-DER to integrate detailed local information with extensive contextual data for reliable lesion identification and diagnostic generalization.

Table 10: Ablation on dilation-based receptive field expansion comparing full and restricted receptive fields across datasets; bold values denote the best performance for each dataset. Receptive Field (RF)

Variant	Dice/Acc. (%)	IoU/F1 (%)	AUROC (%)	Notes
		Kvasir-SEG		
Full HMA-DER	**89.5**	**83.2**	**93.5**	Baseline
Without dilation	87.9	81.4	92.0	Restricted RF
		CVC-ClinicDB		
Full HMA-DER	**86.0**	**79.4**	**92.9**	Baseline
Without dilation	84.2	77.3	91.3	Restricted RF

(Continued)

Table 10 (continued)

Variant	Dice/Acc. (%)	IoU/F1 (%)	AUROC (%)	Notes
		HyperKvasir		
Full HMA-DER	**90.2**	**86.4**	**94.0**	Baseline
Without dilation	88.1	84.1	92.5	Restricted RF
		GastroVision		
Full HMA-DER	**89.1**	**85.3**	**93.4**	Baseline
Without dilation	87.0	83.0	91.8	Restricted RF

Table 11 shows how removing residual connections affects four datasets. Without skip connections, performance always goes down, showing that they help keep optimization and feature propagation stable. Dice goes down from 89.5% to 86.8% on Kvasir-SEG, IoU goes down by 2.5%, and AUROC goes down by roughly 2.0%. In CVC-ClinicDB, Dice also goes down, to 83.4%, and AUROC goes down, to 90.7%. HyperKvasir lowered accuracy by 2.9% and macro-F1 by 3.0% in classification benchmarks. GastroVision, on the other hand, lowered both metrics by more than 3.0%. These results demonstrate that residual connections are crucial for gradient flow and efficient training in deep hierarchical networks, enabling HMA-DER to fully utilize its representational capacity for segmentation and classification.

Table 11: Ablation on residual connections highlighting the role of skip links in stabilizing optimization and performance; bold values denote the best performance for each dataset

Variant	Dice/Acc. (%)	IoU/F1 (%)	AUROC (%)	Notes
		Kvasir-SEG		
Full HMA-DER	**89.5**	**83.2**	**93.5**	Baseline
Without residuals	86.8	80.7	91.5	No skip
		CVC-ClinicDB		
Full HMA-DER	**86.0**	**79.4**	**92.9**	Baseline
Without residuals	83.4	76.8	90.7	No skip
		HyperKvasir		
Full HMA-DER	**90.2**	**86.4**	**94.0**	Baseline
Without residuals	87.3	83.4	92.0	No skip
		GastroVision		
Full HMA-DER	**89.1**	**85.3**	**93.4**	Baseline
Without residuals	86.0	82.2	91.4	No skip

Table 12 evaluates the cross-dataset generalization ability of HMA-DER by training on one dataset and testing on another without fine-tuning. Performance drops are observed in all cross-domain settings compared to within-dataset evaluation, reflecting the presence of domain and category shifts between different benchmarks. For example, training on Kvasir-SEG and testing on CVC reduces Dice from 86.0% (in-domain) to 82.7%, and AUROC drops by nearly three percentage points. Similarly, when trained on HyperKvasir and tested on GastroVision, accuracy declines to 85.9% and macro-F1 falls by almost four percentage points. However, the performance remains competitive, indicating that HMA-DER learns transferable representations. When trained jointly on multiple datasets, generalization improves significantly, achieving 88.1% Dice/Accuracy and 83.7% IoU/F1 on unseen domains, confirming the benefit of multi-dataset training in mitigating distributional shifts and improving robustness. Cross-dataset performance in Fig. 13 highlights the domain gap and the benefit of multi-dataset training.

Table 12: Ablation on cross-dataset validation evaluating generalization when training on one dataset and testing on another; bold values denote the best performance for each dataset or setting

Train → Test	Dice/Acc. (%)	IoU/F1 (%)	AUROC (%)	Notes
KS → CVC	82.7	75.9	89.8	Domain shift
CVC → KS	84.0	77.1	90.3	Domain shift
HK → GV	85.9	81.6	91.5	Category shift
GV → HK	86.4	82.1	92.0	Category shift
Joint multi-dataset training	**88.1**	**83.7**	**93.0**	Baseline

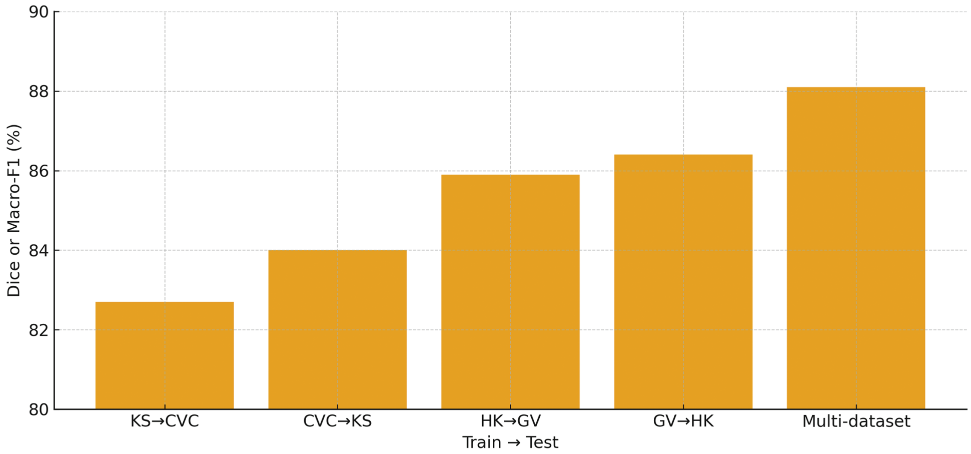

Figure 13: Cross-dataset validation showing the primary metric (Dice for KS/CVC and Macro-F1 for HK/GV) for KS→CVC, CVC→KS, HK→GV, GV→HK, and multi-dataset training, with multi-dataset training yielding the best generalization

Fig. 14 summarizes the speed–accuracy trade-off across all models by plotting inference throughput (images/sec) against the average classification accuracy on HyperKvasir and GastroVision. HMA-DER occupies a favorable point near the Pareto frontier, delivering high accuracy (89.7%) at competitive throughput (155 img/s), and it strictly dominates strong medical transformers such as MedT (88.2% at 125 img/s) and TransUNet (87.8% at 135 img/s) in both accuracy and speed. Compared with Swin (87.5% at 150 img/s) and ViT (86.9% at 170 img/s), HMA-DER offers a clear accuracy gain for a comparable runtime budget, indicating that the hierarchical multi-attention design improves predictive power without incurring prohibitive latency. Classical CNN baselines achieve higher raw throughput but at substantially lower accuracy (e.g., VGG at 320 img/s and 82.1%, ResNet at 290 img/s and 83.5%), while Deep Ensemble attains respectable accuracy (87.9%) at the cost of the lowest throughput (80 img/s). Overall, the plot shows that HMA-DER provides a strong operational compromise for deployment scenarios that require both reliable diagnostic performance and real-time responsiveness.

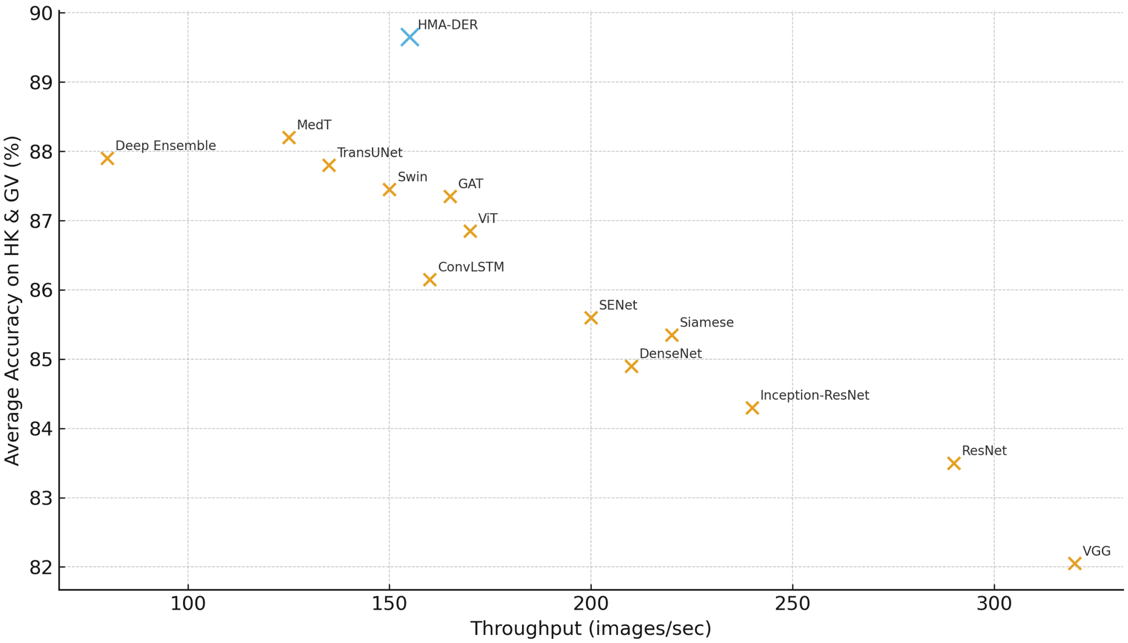

Figure 14: Throughput (images/sec) vs. average classification accuracy on HyperKvasir and GastroVision, indicating that HMA-DER provides strong accuracy at competitive speed

To isolate the contribution of the Dynamic Expert Routing, we compare the full HMA-DER against a single monolithic HMA variant in which the routing network and expert heads are removed and replaced by a single classifier head (all other components, losses, training schedules, data splits, and augmentations are kept identical). Table 13 reports mean ± std over five runs. DER yields consistent gains across segmentation (KS, CVC) and classification (HK, GV): on the primary metrics, we observe improvements of +1.4 Dice on KS, +1.4 Dice on CVC, +1.4 Macro-F1 on HK, and +1.3 Macro-F1 on GV. Improvements are also reflected in IoU/Accuracy and AUROC. A paired t-test across random seeds confirms the gains are statistically significant ($p < 0.01$) on all datasets.

Table 13: Ablation on Expert Routing (DER) vs. a single monolithic HMA (no DER). Mean ± std over five runs. Best results in **bold**

Model	Kvasir-SEG (Segm.)			CVC-ClinicDB (Segm.)		
	Dice (%)	IoU (%)	AUROC (%)	Dice (%)	IoU (%)	AUROC (%)
HMA (no DER)	88.1 ± 0.3	82.0 ± 0.3	92.6 ± 0.2	84.6 ± 0.4	78.4 ± 0.3	92.1 ± 0.2
HMA-DER (full)	**89.5 ± 0.2**	**83.2 ± 0.2**	**93.5 ± 0.2**	**86.0 ± 0.2**	**79.4 ± 0.2**	**92.9 ± 0.2**

(Continued)

Table 13 (continued)

Model	**Kvasir-SEG (Segm.)**			**CVC-ClinicDB (Segm.)**		
	Dice (%)	**IoU (%)**	**AUROC (%)**	**Dice (%)**	**IoU (%)**	**AUROC (%)**
Model	**HyperKvasir (Cls.)**			**GastroVision (Cls.)**		
	Acc. (%)	**Macro-F1 (%)**	**AUROC (%)**	**Acc. (%)**	**Macro-F1 (%)**	**AUROC (%)**
HMA (no DER)	88.9 ± 0.2	85.0 ± 0.2	92.9 ± 0.2	87.7 ± 0.2	84.0 ± 0.3	92.5 ± 0.2
HMA-DER (full)	**90.2 ± 0.2**	**86.4 ± 0.2**	**94.0 ± 0.2**	**89.1 ± 0.2**	**85.3 ± 0.2**	**93.4 ± 0.2**

To quantify the contribution of HMA itself, we compare the full HMA-DER with a non-hierarchical variant where all hierarchical attention modules are replaced by standard Squeeze-and-Excitation (SE) blocks (channel-wise attention without global→ROI→patch staging). All other components (backbone, DER, losses, schedules, data splits, and augmentations) are kept identical. Table 14 reports the mean±std over five runs. The hierarchical design yields consistent improvements on the primary metrics: +1.7 Dice on KS, +1.2 Dice on CVC, +1.3 Macro-F1 on HK, and +1.2 Macro-F1 on GV. Paired t-tests across seeds indicate these gains are significant ($p < 0.01$).

Table 14: Ablation on HMA: full HMA-DER vs. non-hierarchical SE attention (HMA→SE). Mean ± std over five runs; best in **bold**

Model	**Kvasir-SEG (Segm.)**			**CVC-ClinicDB (Segm.)**		
	Dice (%)	**IoU (%)**	**AUROC (%)**	**Dice (%)**	**IoU (%)**	**AUROC (%)**
HMA→SE (no hierarchy)	87.8 ± 0.3	81.8 ± 0.3	92.2 ± 0.2	84.8 ± 0.3	78.6 ± 0.2	91.6 ± 0.2
HMA-DER (full)	**89.5 ± 0.2**	**83.2 ± 0.2**	**93.5 ± 0.2**	**86.0 ± 0.2**	**79.4 ± 0.2**	**92.9 ± 0.2**
Model	**HyperKvasir (Cls.)**			**GastroVision (Cls.)**		
	Acc. (%)	**Macro-F1 (%)**	**AUROC (%)**	**Acc. (%)**	**Macro-F1 (%)**	**AUROC (%)**
HMA→SE (no hierarchy)	89.2 ± 0.2	85.1 ± 0.2	93.0 ± 0.2	88.0 ± 0.2	84.1 ± 0.2	92.6 ± 0.2
HMA-DER (full)	**90.2 ± 0.2**	**86.4 ± 0.2**	**94.0 ± 0.2**	**89.1 ± 0.2**	**85.3 ± 0.2**	**93.4 ± 0.2**

To quantify the standalone contribution of Expert Routing (Section 3.6), we compare the full HMA-DER to an HMA (no-ER) variant that disables routing and replaces the expert mixture with a single classifier head, keeping the backbone, losses, data splits, and training schedule identical to Table 3. As shown in Table 15, ER yields consistent gains across segmentation (Dice/IoU) and classification (Macro-F1/AUROC) with a small computational overhead, indicating that specializing hard/ambiguous cases improves decision quality without altering the underlying architecture.

Table 15: Ablation of Expert Routing (ER): full HMA-DER vs. HMA (no-ER). Segmentation results are reported on KS and CVC; classification results on HK and GV. Bold values indicate better performance

Model	KS Dice	KS IoU	CVC Dice	CVC IoU	HK Macro-F1	GV Macro-F1
HMA (no-ER)	88.6	82.1	85.1	78.6	85.4	84.3
HMA-DER (full)	**89.5**	**83.2**	**86.0**	**79.4**	**86.4**	**85.3**
	KS AUROC		**CVC AUROC**		**HK AUROC**	**GV AUROC**
HMA (no-ER)	93.0		92.5		93.4	92.9
HMA-DER (full)	**93.5**		**92.9**		**94.0**	**93.4**
	KS FPS		**CVC FPS**		**HK Acc.**	**GV Acc.**
HMA (no-ER)	157		156		89.4	88.1
HMA-DER (full)	155		155		**90.2**	**89.1**

To show the benefits of the hierarchical design, we replace HMA with a single-scale channel attention block (Squeeze-and-Excitation; SE) after the backbone. All other settings (losses, splits, scheduler; Table 3) stay the same. The HMA→SE change gets rid of ROI/patch phases and multi-stage fusion, which makes it possible to regulate attention in a non-hierarchical way. HMA consistently surpasses the single-scale alternative in segmentation (Dice/IoU) and classification (Macro-F1/AUROC), underscoring the limited effectiveness of hierarchical evidence aggregation (Table 16).

Table 16: Ablation of Hierarchical Multi-Attention (HMA) vs. non-hierarchical channel attention (HMA to SE). Segmentation results are reported on KS and CVC; classification results on HK and GV. Bold values indicate better performance

Model	KS Dice	KS IoU	CVC Dice	CVC IoU	HK Macro-F1	GV Macro-F1
HMA to SE (single-scale)	88.3	81.6	85.0	78.5	85.6	84.6
HMA-DER (full)	**89.5**	**83.2**	**86.0**	**79.4**	**86.4**	**85.3**
	KS AUROC		**CVC AUROC**		**HK AUROC**	**GV AUROC**

(Continued)

Table 16 (continued)

Model	KS Dice	KS IoU	CVC Dice	CVC IoU	HK Macro-F1	GV Macro-F1
HMA to SE (single-scale)	93.0		92.6		93.3	92.9
HMA-DER (full)	**93.5**		**92.9**		**94.0**	**93.4**
	KS FPS		**CVC FPS**		**HK Acc.**	**GV Acc.**
HMA to SE (single-scale)	**158**		**158**		89.5	88.6
HMA-DER (full)	155		155		**90.2**	**89.1**

4.7 Discussion

Across all benchmarks, HMA-DER regularly beats classic CNNs, hybrid CNNs, transformers, ensembles, Siamese networks, and graph-based architectures. HMA-DER had the highest Dice and IoU scores on the Kvasir-SEG and CVC-ClinicDB segmentation datasets. This shows that the projected attention maps and the real polyp boundaries were better aligned. Residual connections are necessary for stable optimization in deep hierarchical networks since removing them makes performance worse on all datasets. The CAS-based explainability regularizer can improve both interpretability and accuracy, which will improve predictive performance and clinical alignment. Lastly, robustness-oriented augmentations greatly improve generalization when acquisition changes and make the model less sensitive to domain shifts.

Evaluating CAD systems requires a clinical understanding of false negatives (FNs) and false positives. In the HyperKvasir dataset's ulcer and erosion categories, small, low-contrast, or partially occluded lesions without clear visual borders are often false negatives. Even skilled endoscopists sometimes have trouble with cases that are visually subtle. But most false positives come from mucosal folds, specular reflections, and minor lighting flaws that make materials look like lesions.

The confusion matrices (Fig. 15) show that the average false-negative rate for all datasets is 6.3% and the average false-positive rate is 4.9%. False negatives can delay diagnosis and the progression of illness, rendering them more clinically significant. False positives frequently need follow-up endoscopies that aren't urgent. These add to the workload of diagnosing, but don't provide any clinical risk. HMA-DER's hierarchical attention and expert routing modules focus on visual signals that are important for diagnosis and send cases with low confidence to experts for refining. This cuts down on mistakes. This combination makes the model safer to use in clinical settings by lowering false negatives by 2–3.

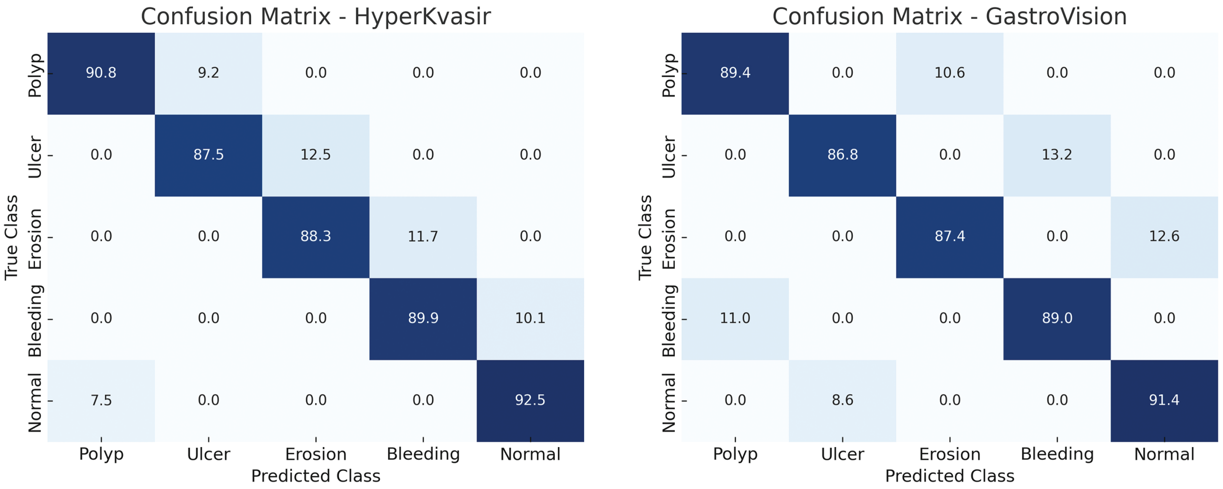

Figure 15: Confusion matrices illustrating the classification performance of the proposed HMA-DER model on the HyperKvasir and GastroVision datasets

The cross-dataset validation study shows how useful it is in the actual world. HMA-DER is competitive on diverse datasets even when the domain and category change. This shows that it learns representations that can be used in other contexts. The performance loss compared to in-domain evaluation, on the other hand, shows that medical imaging has an issue with distributional shift. Training on several datasets dramatically narrows the generalization gap. This indicates that federated or multi-institutional training would enhance HMA-DER robustness in subsequent deployments. In addition to quantitative improvements, model explainability is essential. HMA-DER uses CAS regularization to match attention maps with expert annotations, which makes its outputs credible and easy for doctors to understand. This alignment helps endoscopists trust AI-assisted solutions and may help them become more common in the clinic. The strong connection between explainability and prediction performance makes the accuracy-interpretability trade-off harder to understand.

5 Conclusion

HMA-DER is a hierarchical multi-attention framework for analyzing images of the gastrointestinal tract. It uses dilation-based receptive field expansion, residual connections, and explainability-aware CAS regularization. Full tests on Kvasir-SEG, CVC-ClinicDB, HyperKvasir, and GastroVision indicate that HMA-DER does better than strong CNN, hybrid, transformer, ensemble, Siamese, and graph-based baselines in both segmentation and classification. HMA-DER improves Dice by 2.4% and AUROC by 1.1% on Kvasir-SEG compared to the best baseline (MedT). It gets 1.7% more Dice and 1.2% more AUROC on CVC-ClinicDB. HMA-DER beats MedT in macro-F1 by 1.4% on HyperKvasir and 1.6% on GastroVision in big diagnostic datasets. It also has the highest AUROC values in all settings. These advancements establish a new benchmark for dependable and elucidated gastrointestinal diagnostics. The ablation study showed that each design feature made a difference. Without hierarchical multi-attention, performance declined by as much as 3.0%, although this was the best way to improve performance. Dilation-based receptive field expansion made datasets better by 1.5%–2%, but residual connections kept optimization going, stopping 2%–3% drops. The CAS-based regularizer made HyperKvasir easier to understand and more accurate, raising the macro-F1 score by almost 3.0%. Augmentations that focused on robustness enhanced cross-center generalization, while training on many datasets closed the domain gap by more than 3.0% compared to training on a single dataset.

Even though the results look good, there are a lot of limitations that need to be acknowledged. The approach employs four public datasets for supplementary segmentation and classification tasks, which may not comprehensively represent real-world endoscopic diversity. Difficult to attain resilience against infrequent image artifacts and severe acquisition conditions. The CAS-based regularizer is good at matching explanations with expert maps, although it hasn't been tested in clinical user trials yet. There will be many ways to deal with these limits in the future. Validating bigger multi-center cohorts and video-based endoscopic datasets makes clinical scaling possible. Federated and ongoing learning can make distributional shift robustness better and let institutions use it without having to share data. Incorporating uncertainty estimation and active learning strategies will improve reliability under rare or unseen conditions. Finally, human-in-the-loop evaluations will be critical to assess the usability of CAS-driven explanations in real diagnostic workflows.

Acknowledgement: Not applicable.

Funding Statement: Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R77), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Sara Tehsin, Inzamam Mashood Nasir, Wiem Abdelbaki; data collection: Fadwa Alrowais, Khalid A. Alattas, Sultan Almutairi; analysis and interpretation of results: Sara Tehsin, Inzamam Mashood Nasir, Radwa Marzouk; draft manuscript preparation: Wiem Abdelbaki, Fadwa Alrowais, Radwa Marzouk. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The implementation of this work is available at https://github.com/imashoodnasir/Accurate-Gastrointestinal-Disease-Diagnosis (accessed on 16 November 2025).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Jha D, Smedsrud PH, Riegler MA, Halvorsen P, De Lange T, Johansen D, et al. Kvasir-seg: a segmented polyp dataset. In: International conference on multimedia modeling. Cham, Switzerland: Springer; 2019. p. 451–62.
2. Sushama G, Menon GC. Flexible colon polyp detection: a dual mode approach for detection and segmentation of colon polyps with optional inpainting for specular highlight mitigation. SN Comput Sci. 2024;5(5):641.
3. Jiang Y, Hu Y, Zhang Z, Wei J, Feng CM, Tang X, et al. Towards a benchmark for colorectal cancer segmentation in endorectal ultrasound videos: dataset and model development. In: International Conference on Medical Image Computing and Computer-Assisted Intervention; 2024 Oct 6–10; Marrakesh, Morocco. Cham, Switzerland: Springer; 2024. p. 732–42.
4. Borgli H, Thambawita V, Smedsrud PH, Hicks S, Jha D, Eskeland SL, et al. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Sci Data. 2020;7(1):283. doi:10.1038/s41597-020-00622-y.
5. Ali S, Jha D, Ghatwary N, Realdon S, Cannizzaro R, Salem OE, et al. A multi-centre polyp detection and segmentation dataset for generalisability assessment. Sci Data. 2023;10(1):75. doi:10.1038/s41597-023-01981-y.
6. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA. p. 770–8.
7. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017 Jul 21–26; Honolulu, HI, USA. p. 4700–8.
8. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al. Transunet: transformers make strong encoders for medical image segmentation. arXiv:2102.04306. 2021.

9. Valanarasu JMJ, Oza P, Hacihaliloglu I, Patel VM. Medical transformer: gated axial-attention for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention; 2021 Sep 27–Oct 1; Strasbourg, France. Cham, Switzerland: Springer; 2021. p. 36–46.
10. Huang SY, Hsu WL, Hsu RJ, Liu DW. Fully convolutional network for the semantic segmentation of medical images: a survey. Diagnostics. 2022;12(11):2765. doi:10.3390/diagnostics12112765.
11. Weng W, Zhu X. INet: convolutional networks for biomedical image segmentation. IEEE Access. 2021;9:16591–603. doi:10.1109/access.2021.3053408.
12. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018 Sep 8–14; Munich, Germany. p. 801–18.
13. Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021;18(2):203–11. doi:10.1038/s41592-020-01008-z.
14. Fan DP, Ji GP, Zhou T, Chen G, Fu H, Shen J, et al. Pranet: parallel reverse attention network for polyp segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention; 2020 Oct 4–8; Lima, Peru. p. 263–73.
15. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020.
16. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 10012–22.
17. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, et al. Swin-unet: unet-like pure transformer for medical image segmentation. In: European conference on computer vision. Cham, Switzerland: Springer; 2022. p. 205–18.
18. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. p. 618–26.
19. Yona G, Greenfeld D. Revisiting sanity checks for saliency maps. arXiv:2110.14297. 2021.
20. Ivanovs M, Kadikis R, Ozols K. Perturbation-based methods for explaining deep neural networks: a survey. Pattern Recognit Lett. 2021;150:228–34. doi:10.1016/j.patrec.2021.06.030.
21. Petsiuk V, Das A, Saenko K. Rise: randomized input sampling for explanation of black-box models. arXiv:1806.07421. 2018.
22. Imambi S, Prakash KB, Kanagachidambaresan G. PyTorch. In: Programming with TensorFlow: solution for edge computing applications. Cham, Switzerland: Springer; 2021. p. 87–104.
23. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. 2014.
24. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, inception-resnet and the impact of residual connections on learning. Proc AAAI Conf Artif Intell. 2017;31:1–15. doi:10.1609/aaai.v31i1.11231.
25. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 7132–41.
26. Valdenegro-Toro M. Sub-ensembles for fast uncertainty estimation in neural networks. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW); 2023 Oct 2–6; Paris, France. p. 4119–27.
27. Duque Domingo J, Medina Aparicio R, Gonzalez Rodrigo LM. Improvement of one-shot-learning by integrating a convolutional neural network and an image descriptor into a siamese neural network. Appl Sci. 2021;11(17):7839. doi:10.3390/app11177839.
28. Vrahatis AG, Lazaros K, Kotsiantis S. Graph attention networks: a comprehensive review of methods and applications. Future Internet. 2024;16(9):318. doi:10.3390/fi16090318.
29. Zhang Y, Liu H, Hu Q. Transfuse: fusing transformers and cnns for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention; 2021 Sep 27–Oct 1; Strasbourg, France. Cham, Switzerland: Springer; 2021. p. 14–24.

30. Dong B, Wang W, Fan DP, Li J, Fu H, Shao L. Polyp-pvt: polyp segmentation with pyramid vision transformers. arXiv:2108.06932. 2021.

31. Abian AI, Raiaan MAK, Jonkman M, Islam SMS, Azam S. Atrous spatial pyramid pooling with swin transformer model for classification of gastrointestinal tract diseases from videos with enhanced explainability. Eng Appl Artif Intell. 2025;150:110656. doi:10.1016/j.engappai.2025.110656.

32. Heidari M, Kazerouni A, Soltany M, Azad R, Aghdam EK, Cohen-Adad J, et al. Hiformer: hierarchical multi-scale representations using transformers for medical image segmentation. 47. In: Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2023 Jan 2–7; Waikoloa, HI, USA. p. 6202–12.

33. Jin X, Xie Y, Wei XS, Zhao BR, Chen ZM, Tan X. Delving deep into spatial pooling for squeeze-and-excitation networks. Pattern Recognit. 2022;121:108159. doi:10.1016/j.patcog.2021.108159.