MoE-Polyp: Shifting More Attention to Small Polyp Segmentation via Mixture-of-Experts

Zihuang Wu
School of Computer and Information Engineering, Jiangxi Normal University
China
1874wzh@gmail.com

Xinyu Xiong
Hangzhou Hikvision Digital Technology Co Ltd
China
School of Computer Science and Engineering, Sun Yat-sen University
China
xiongxyowo@gmail.com

Ying Chen
Pazhou Lab
China
chenying980812@gmail.com

Siying Li
Smart Hospital Research Institute, Peking University Shenzhen Hospital
China
2070246077@email.szu.edu.cn

Hua Chen*
School of Computer and Information Engineering, Jiangxi Normal University
China
chenhua5752@hotmail.com

Abstract

Accurate segmentation of various types of polyps from colonoscopic images plays an increasingly important role in the computer-assisted diagnosis of colon cancer. Despite recent advances, the identification of small polyps remains a challenging task. In this paper, we find that the existing U-shaped framework is prone to losing valuable features during the cascade decoding process. To address this issue, we propose a mixture-of-experts style segmentation decoder. Our framework, equipped with boundary, spatial, and global experts, adaptively selects the most appropriate features to predict polyps of various sizes. Additionally, we introduce a new benchmark specifically designed to better evaluate the performance on small polyps. Extensive experiments on this benchmark demonstrate that MoE-Polyp outperforms other methods on both standard and small polyps. The established new benchmark is available at https://github.com/WZH0120/MoE-Polyp.

CCS Concepts

• **Computing methodologies** → **Image segmentation**; • **Applied computing** → **Health informatics**.

Keywords

Polyp segmentation, medical image analysis, mixture-of-experts, small object segmentation.

*Corresponding author.

MMASIA '24, December 03–06, 2024, Auckland, New Zealand

ACM ISBN 979-8-4007-1273-9/24/12
https://doi.org/10.1145/3696409.3700169

ACM Reference Format:
Zihuang Wu, Xinyu Xiong, Ying Chen, Siying Li, and Hua Chen. 2024. MoE-Polyp: Shifting More Attention to Small Polyp Segmentation via Mixture-of-Experts. In *ACM Multimedia Asia (MMASIA '24), December 03–06, 2024, Auckland, New Zealand*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3696409.3700169

1 Introduction

Medical image segmentation [2, 15, 18, 23, 26] plays an important role in computer-aided diagnosis. Accurately identifying polyps from colonoscopy images [6, 12, 27] can help clinicians effectively locate and segment polyps, improving their efficiency and reducing the miss rate. As colorectal cancer ranks high in both incidence and mortality among major malignancies worldwide [22], extensive deep-learning methods have been developed to assist in colorectal cancer screening, early diagnosis, and treatment, thereby effectively increasing the survival rate. However, despite the progress of these approaches, performance degradation on small polyps [11, 30] remains a significant challenge. Small polyps often resemble the background normal tissue in texture and color, making them difficult to identify accurately.

To address this issue, existing approaches primarily focus on designing advanced attention mechanisms. However, these methods are generally based on the U-shaped structure, as illustrated in Figure 1(a). Originating with U-Net [20], this structure has demonstrated excellent versatility in many medical image segmentation tasks. Specifically, the decoder of the U-shaped network operates in a cascade manner, gradually restoring decoding features from the deep layer to the shallow layer to produce the final output. During this process, features at different levels are concatenated, which can result in the loss of valuable information during successive propagation through the decoder, thereby compromising segmentation performance. Additionally, the gradual concatenation of features leads to redundancy and increased computational overhead.

To overcome the above shortcomings of the existing U-decoder, we seek inspiration from the emerging mixture-of-experts (MoE) [21]

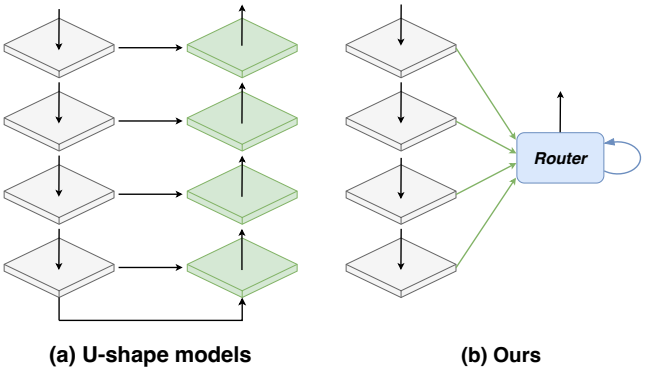

Figure 1: Conception comparasion between existing U-Net [20] like segmentation methods [6, 8, 24, 31] and our MoE-Polyp. The green part denotes leveraging attention mechanism to enhance features and the blue part denotes the router (also known as gating network) in MoE design.

design in nature language processing (NLP). Originally, the MoE is proposed to allow efficient model scaling with a limited increase in computational cost. Specifically, it consists of many experts, each of which is a learnable neural network, along with a trainable gating network that selects a suitable combination of experts to process each input.

Based on these observations, we abandoned the traditional U-shaped structure and proposed a polyp segmentation method based on the mixture-of-experts design, as illustrated in Figure 1(b). Specifically, we process the features from the encoder through different experts and learn a gating network to select a suitable set of expert features to predict the label for each pixel. By using MoE, the model can learn more specialized and disentangled expert feature maps and reduce interference between them during inference. Furthermore, the decoder enables better specialization and disentanglement of feature maps, which substantially improve performance.

In summary, our main contributions are as follows:

- A novel polyp segmentation framework named MoE-Polyp that able to adaptively emphasize the appropriate features according to the size of the polyp to obtain better results.
- A new polyp segmentation benchmark to improve research on small polyp segmentation within the community.

2 Related Work

2.1 Polyp Segmentation

With the significant advancements in deep learning, many convolutional neural network-based methods have been applied to image segmentation. Among them, U-Net [20] has gained widespread attention for its effectiveness in medical image segmentation. Inspired by the success of U-Net in medical image segmentation, many variants of U-Net have been developed for medical image segmentation. For example, PraNet [6] proposes using a reverse attention module to exploit boundary cues, thereby establishing the relationship between boundaries and regions. SANet [24] designs a shallow attention mechanism to filter background noise from

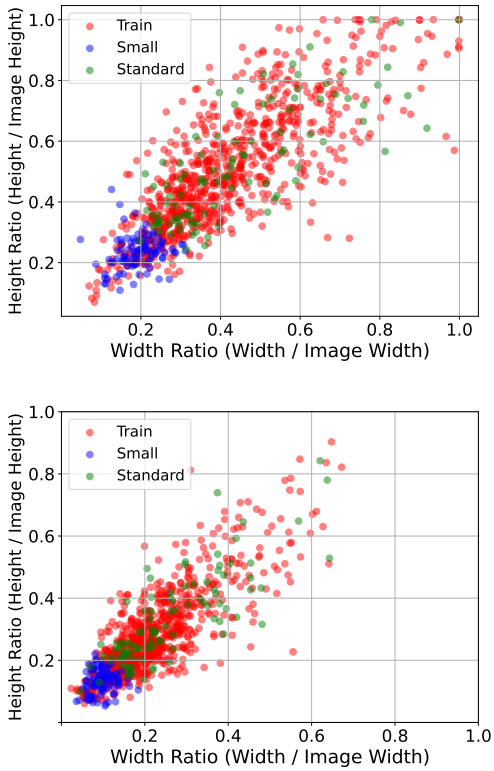

Figure 2: Data distribution statics of Kvasir-SEG (top) and BKAI-IGH (down) dataset.

shallow features. CaraNet [17] introduced an axial reverse attention module to analyze target position information and multi-scale features, enhancing sensitivity to significant information.

2.2 Mixture-of-Experts

In computer vision, MoE is also used in many tasks. For example, the classification network in Norface [13] relies on MoE to enhance feature representation for facial expression analysis. MoVA [35] uses mixed visual experts and expert routing to extract and fuse task-specific knowledge from various experts, enhancing the model's generalization ability. In [16], an MoE structure is used to increase learnable parameters to model the multifaceted nature of natural and generated images. MLoRE [28] adds a simple universal convolutional path to MoE, allowing different tasks to share it and alleviating global relation modeling. Different from existing MoE strategies that are generally adopted in conjunction with large-scale foundation models [14] for parameter-efficient fine-tuning, in this paper, we reveal that MoE can also be a good segmentation decoder.

3 The New Benchmark

Existing benchmarks [6] for evaluating polyp segmentation performance fails to account for size differences among polyps, treating all sizes uniformly. However, some studies [17] have shown that a major bottleneck in current polyp segmentation algorithms is the

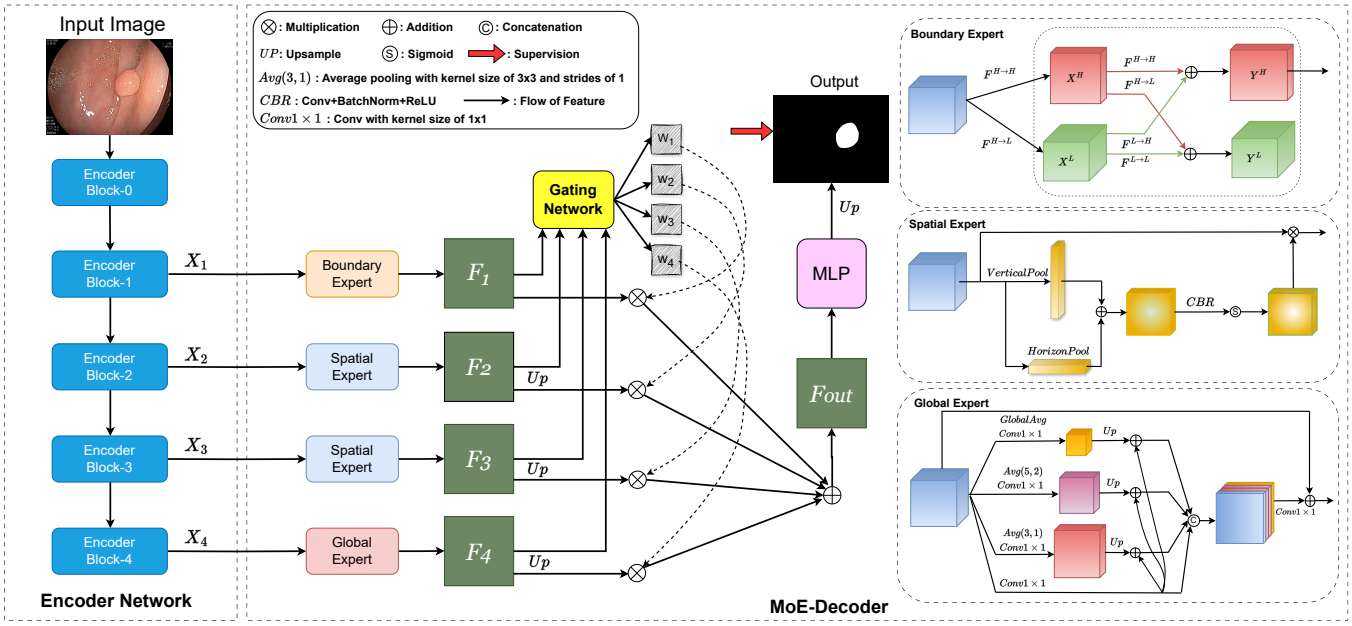

Figure 3: Overview of our proposed MoE-Polyp framework.

detection of small polyps, and existing evaluation frameworks overlook the need for separate analysis of these challenging samples.

To obtain a more comprehensive understanding of polyp size distribution, we dive into a closer look at two well-known public polyp segmentation datasets: Kvasir-SEG (contains 1000 polyp images) [10] and BKAI-IGH (contains 1000 polyp images) [19] datasets, as depicted in Figure 2. From this figure, we can observe that both datasets contain a significant number of small polyp samples.

To encourage advancements in the performance of algorithms on small polyps, we reclassified these two datasets. Specifically, we sorted the polyp images according to their polyp size. From the smallest 20% of the samples, we randomly selected 10% as the "small" subset of our test set, enabling independent performance evaluation on small polyps. Additionally, from the remaining 80% of the samples, which represent more typical sizes, we randomly selected 10% to constitute the "standard" subset of our test set. Apart from the small and standard test sets, the rest are used as the training and validation sets.

4 Method

4.1 Framework Overview

In this paper, we propose a mixture-of-experts-based polyp segmentation network (MoE-Polyp), as illustrated in Figure 3. The network includes a feature extraction backbone and a MoE-style decoder that contains three different types of experts. Given an input image $I \in \mathbb{R}^{3\times H\times W}$, where H, and W denote the height and width respectively, it is first fed into the Res2Net [7] backbone to obtain multi-level features $X_i \in \mathbb{R}^{C_i\times \frac{H}{2^{i+1}}\times \frac{W}{2^{i+1}}}$ ($i \in \{1,2,3,4\}, C_i \in \{256, 512, 1024, 2048\}$), where C denotes number of channels. Since features at different levels contain information of varying granularity, we feed them into different expert modules for specialization. The outputs of these expert modules, denoted as $F_i (i \in 1,2,3,4)$, are then sent to a gating network for reweighting. The reweighted features are finally aggregated and decoded to produce the polyp segmentation result. In the following sections, we will describe the expert modules and the mixture-of-experts design in detail.

4.2 Boundary Expert

Due to the high similarity between certain polyps and the surrounding normal tissue, identifying the boundaries of polyps is challenging. Existing methods primarily focus on designing additional supervision in the RGB domain to improve identification accuracy [9, 29]. However, these methods often overlook valuable information present in the frequency domain. In an image, low-frequency features correspond to gradient intensity transitions in pixel regions, such as large color blocks, typically representing the main parts of objects. In contrast, high-frequency components involve pixels with significant brightness changes, such as object boundaries. Compared to RGB appearance features, frequency information has significant advantages in identifying camouflaged objects [33].

Inspired by the above observations, we designed a frequency-domain guided boundary expert. The goal of our boundary expert is to effectively utilize high-frequency information to better perceive the boundaries of polyps. Specifically, since low-level features have a large resolution and contain richer boundary information compared to high-level features, we use octave convolution [4] to automatically perceive high-frequency and low-frequency information. The detailed process of octave convolution output can be described as follows:

$$Y^H = F(X^H; W^{H\to H}) + Upsmple(F(X^L; W^{L\to H}), 2),$$
$$Y^L = F(X^L; W^{L\to L}) + F(pool(X^H, 2); W^{H\to L}), \tag{1}$$

where F(X; W) denotes a convolution with the learnable parameters of W, pool(X, K) is an average pooling operation with a kernel size of k × k, and Upsample(X, n) is an up-sampling operation by a factor of n via nearest interpolation.

Then, we discard low-frequency information and retain high-frequency boundary information:

$$F_1 = Y^H. \tag{2}$$

4.3 Spatial Expert

The spatial expert aims to retain more spatial details about both the interior and exterior of the polyp. We employ horizontal pooling and vertical pooling to capture axial global context in both directions, and use broadcast addition to model rectangular regions of interest, enabling the model to focus more on the foreground polyp areas. Additionally, the spatial expert uses intermediate layer features with higher resolution as input, which helps to preserve rich spatial details. Specifically, given the input middle-level feature $X_i (i \in 2, 3)$, the spatial expert first leverages horizontal pooling $\mathcal{H}(\cdot)$ and vertical pooling $\mathcal{V}(\cdot)$ to capture the axial global context in two directions:

$$D_h = \mathcal{H}(X_i),$$
$$D_v = \mathcal{V}(X_i), \tag{3}$$

where D_h and D_v represents horizontal and axial contexts, respectively. Then, these two contexts with different directions are broadcasted and added to obtain a rectangle region of interest that denotes the coarse possible location of the polyp. Finally, the obtained coarse region are sent into a 3×3 convolution to get refined and get the spatial attention map D_s:

$$D_s = Conv_{3\times3}(D_h \oplus D_v). \tag{4}$$

The obtained attention maps are multiplied with the original input feature X_i and obtained the spatial expert output $F_i (i \in 2, 3)$:

$$F_i = X_i \otimes D_s, \tag{5}$$

where $\otimes$ denotes element-wise multiplication.

4.4 Global Expert

Global and multi-scale features are crucial for image segmentation. However, the original receptive field of the output of the encoder network is limited. Inspired by ASPP [3], PSP [32], and their variants, which use different branches to extract multi-scale features, we design a global expert to capture rich semantic information and multi-scale context with more effective receptive field. Specifically, given a high-level feature $X_i (i = 4)$, it is separately passed through different pooling layers and convolution layers with different kernel sizes and strides to obtain multi-scale features D_i:

$$D_1 = GAP(X_i),$$
$$D_2 = Avg_{(5,2)}(X_i),$$
$$D_3 = Avg_{(3,1)}(X_i),$$
$$D_4 = Conv_{1\times1}(X_i), \tag{6}$$

where GAP denotes global average pooling and Avg denotes average pooling. Then, we leverage D_4 to further guide the features of other scales $D_i (i \in 1, 2, 3)$, which allows the efficient interaction between different features:

$$D_i = Conv_{3\times3}(\uparrow (D_i) \oplus D_4), \tag{7}$$

where $\uparrow$ denotes the upsample operation. Finally, the multi-scale intermediate features D_i are concatenated and sent to a 1×1 conv to reduce the channels. The obtained features are added with the input feature, which obtains the output of global expert:

$$F_4 = Conv_{1\times1}(Cat(D_1, D_2, D_3, D_4)) \oplus X_i, \tag{8}$$

4.5 Expert Routing

After obtaining the outputs of different experts $F_i (i \in 1, 2, 3, 4)$, a vanilla strategy is to directly fuse them by adding or concatenation and then send them to the decoder head. However, since different expert features are derived from experts with different preferences, fusing them directly can cause feature interference and impair segmentation performance.

Inspired by the recent success of the Mixture-of-Experts (MoE) design, we propose a Moe-style decoder that can adaptively select the appropriate parts of each expert feature to participate in the fusion process. Our goal is to obtain four weight maps $[W_1, W_2, W_3, W_4]$ to reweight the upsampled expert features, where each weight map is of size $\frac{H}{2} \times \frac{W}{2}$. Note that the weight maps are constrained by $W_1 + W_2 + W_3 + W_4 = \mathbf{1}$.

Specifically, these weight maps are produced by a gateing network. The gating network first concatenates all the expert feature maps along channels and uses a MLP layer and a final softmax layer to process the concatenated features into the weight maps. We then use the weight maps to produce the combined feature map F_{out}:

$$F_{out} = \sum_{i=1}^{4} W_i \odot F_i, \tag{9}$$

where $\odot$ denotes pixel-wise multiplication. Then, the combined feature map F_{out} is further fused through a MLP, and a convolution with a kernel size of 1 is used to produce the final segmentation result O:

$$O = Conv1 \times 1(MLP(F_{out})). \tag{10}$$

The MoE design of the decoder allows the network to learn more specialized feature maps and reduce the interference between them. For the prediction of each pixel, the gating function chooses a suitable set of features by weighing the importance of global vs. local features.

4.6 Loss Function

We use a combination of BCE loss and IoU loss as our loss function, and its effectiveness has been demonstrated in [6, 25]. Therefore, our loss function can be expressed as: $L = L^w_{BCE} + L^w_{IoU}$, where L^w_{BCE} and L^w_{IoU} represent weighted binary cross entropy (BCE) loss and weighted intersection over union (IoU) loss, respectively. Unlike standard BCE loss and standard IoU loss, weighted BCE loss and weighted IoU loss focus more on difficult pixels rather than assigning equal weight to all pixels.

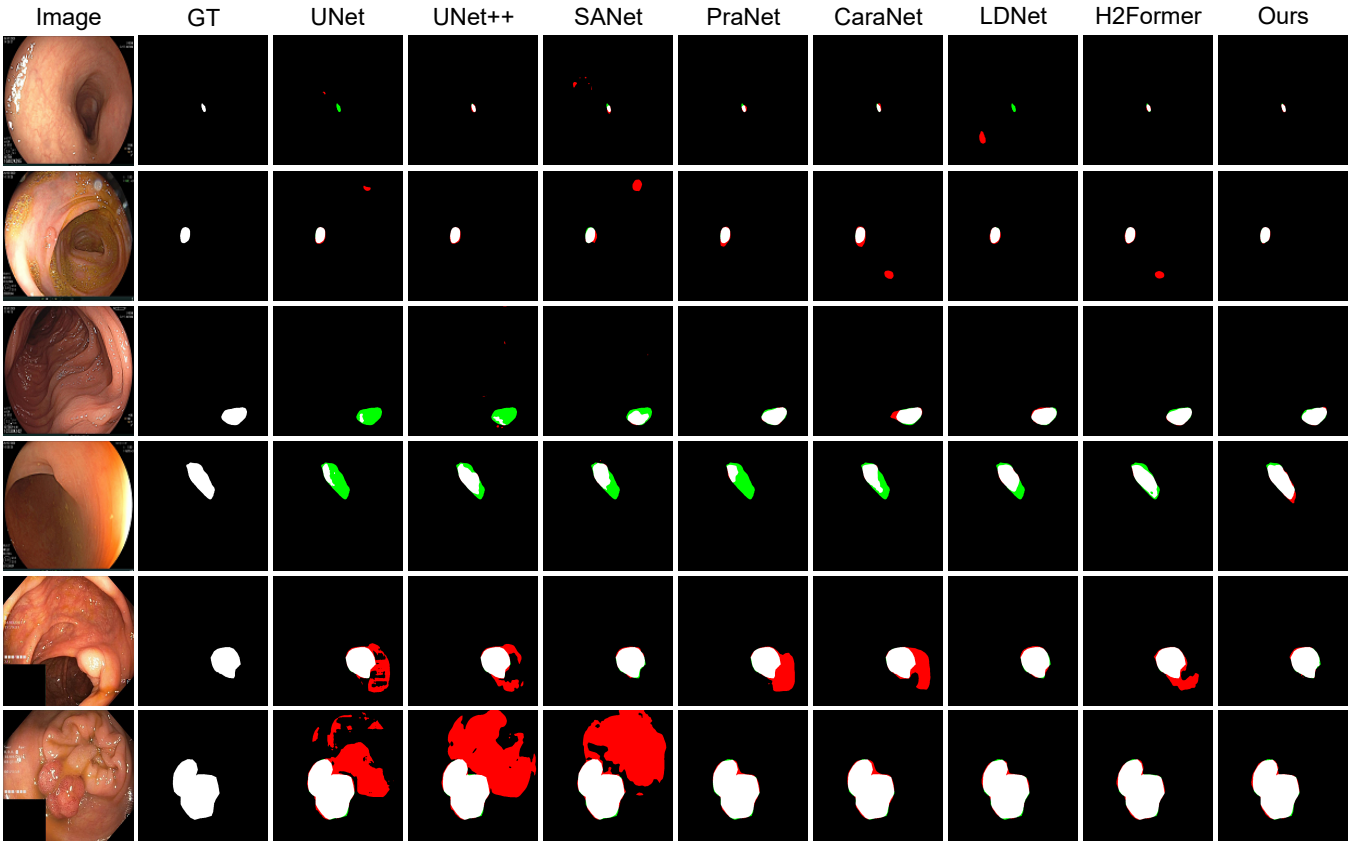

Figure 4: Visual comparison results on polyp segmentation. The white, black, green, and red colors in each prediction map represent the true positive, true negative, false negative, and false positive regions, respectively.

5 Experiments

5.1 Experimental Setup

5.1.1 Datasets. As demonstrated in Section 3, our experiments are conducted on the new benchmark of Kvasir-SEG [10] and BKAI-IGH [19], which can more comprehensively evaluate polyp segmentation performance under various scenes.

5.1.2 Implementation Details. We implement our MoE-Polyp with the PyTorch framework and use an NVIDIA RTX 3090 to accelerate the calculations. We resize all the input images to 352 × 352 for the training and employ a multi-scale training strategy {0.75, 1, 1.25}. We employ the Adam optimization algorithm to optimize the overall parameters with a learning rate of $1e-4$. For data augmentation, we employ random rotation, and random flipping on both horizontal and vertical. We train our network with a batch size of 16 for over 50 epochs.

5.1.3 Compared Methods. Our proposed MoE-Polyp is compared with the following state-of-the-art segmentation methods, including U-Net [20], U-Net++ [34], PraNet [6], SANet [24], LDNet [31], CaraNet [17], and H2Former [8].

5.1.4 Evaluation Metrics. We leverage four widely used metrics to evaluate polyp segmentation performance, including mean dice (mDice), mean intersection over union (mIoU), S-measure (S_α) [5], and weighted F-measure (F_β^ω) [1].

5.2 Ablation Study

In this section, we analyze the effectiveness of each expert module through ablation experiments. All experiments are conducted under the same experimental settings.

5.2.1 Effectiveness of Boundary Expert. The boundary expert aims to provide the model with high-frequency boundary information, enabling more accurate segmentation boundaries. We remove the boundary expert while keeping the other modules unchanged. As shown in the first row of Table 1, the model's performance declined after the removal of the boundary expert. Specifically, the mDice decreased by 3.33%, 0.14%, 1.59%, and 1.49% on Kvasir-Small, Kvasir-Standard, BKAI-Small, and BKAI-Standard, respectively. These results indicate that the boundary expert is crucial for polyp segmentation, particularly for small polyp segmentation.

5.2.2 Effectiveness of Spatial Expert. The spatial expert primarily enriches the model with detailed spatial information. As shown in the second row of Table 1, the model's performance decreased across all four datasets after the spatial expert is removed. Specifically, in terms of mDice, the decreases were 3.11%, 0.87%, 1.04%,

Table 1: Ablation results on Kvasir-SEG and BKAI-IGH.

Dataset	Kvasir-Small				Kvasir-Standard				BKAI-Small				BKAI-Standard			
	mDice	mIoU	S_α	F_β^ω	mDice	mIoU	S_α	F_β^ω	mDice	mIoU	S_α	F_β^ω	mDice	mIoU	S_α	F_β^ω
w/o Boundary Expert	85.33	78.70	90.51	82.93	92.07	86.98	91.59	91.06	89.14	82.67	93.21	87.78	94.22	90.01	95.58	93.77
w/o Spatial Expert	85.55	78.99	90.08	83.07	91.34	86.02	91.03	90.30	89.69	83.62	93.59	88.45	93.05	89.13	94.91	92.09
w/o Global Expert	85.86	79.09	90.71	83.44	91.96	86.51	91.58	90.86	89.75	83.49	93.92	88.12	93.82	89.90	95.17	93.08
MoE-Polyp	**88.66**	**82.66**	**92.40**	**87.13**	**92.21**	**87.19**	**91.95**	**91.58**	**90.73**	**85.39**	**94.46**	**90.12**	**95.71**	**92.48**	**96.51**	**95.57**

Table 2: Results on Kvasir-SEG.

Method	Kvasir-Small				Kvasir-Standard			
	mDice	mIoU	S_α	F_β^ω	mDice	mIoU	S_α	F_β^ω
U-Net	76.17	67.19	84.48	71.49	86.43	78.60	86.57	83.85
UNet++	80.78	73.15	87.59	77.86	86.36	79.54	87.54	83.73
SANet	79.80	71.33	86.72	76.14	87.70	80.43	88.28	86.18
PraNet	87.62	80.92	91.87	85.73	91.28	85.74	91.10	90.04
CaraNet	86.70	79.97	91.15	84.95	91.40	86.17	91.39	90.08
LDNet	85.45	78.23	90.35	83.01	91.39	86.09	91.37	90.24
H2Former	84.80	78.16	89.96	82.37	91.28	86.02	91.12	90.25
Ours	**88.66**	**82.66**	**92.40**	**87.13**	**92.21**	**87.19**	**91.95**	**91.58**

Table 3: Results on BKAI-IGH.

Method	BKAI-Small				BKAI-Standard			
	mDice	mIoU	S_α	F_β^ω	mDice	mIoU	S_α	F_β^ω
U-Net	80.82	72.74	88.98	77.09	87.28	81.40	91.12	85.88
UNet++	81.98	74.34	89.62	78.75	89.90	84.32	92.43	88.10
SANet	80.83	71.85	88.23	78.08	88.01	81.62	91.45	87.35
PraNet	87.30	79.83	92.37	85.08	93.06	88.91	94.60	92.36
CaraNet	88.54	81.64	93.07	87.28	93.68	89.64	95.13	93.26
LDNet	86.10	78.78	91.44	85.45	92.77	87.89	94.46	92.29
H2Former	85.74	78.59	91.39	83.69	92.07	87.63	94.30	90.90
Ours	**90.73**	**85.39**	**94.46**	**90.12**	**95.71**	**92.48**	**96.51**	**95.57**

and 2.66% on Kvasir-Small, Kvasir-Standard, BKAI-Small, and BKAI-Standard, respectively. These results clearly demonstrate the critical role of the spatial expert in enhancing model performance.

5.2.3 Effectiveness of Global Expert. The global expert provides the model with essential global context and multi-scale features. As shown in the third row of Table 1, the model's performance declined after the removal of the global expert. Specifically, in terms of mDice, the decreases were 2.8%, 0.25%, 0.98%, and 1.89% on Kvasir-Small, Kvasir-Standard, BKAI-Small, and BKAI-Standard datasets, respectively. These results indicate the positive impact of the global expert on enhancing the model's segmentation performance.

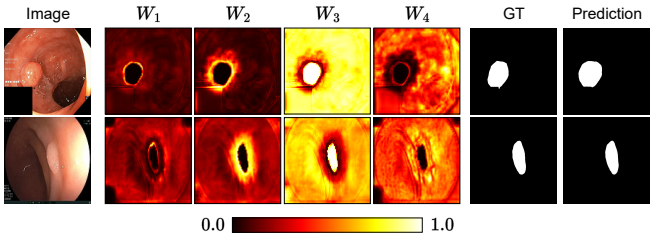

Figure 5: Visualization of four MoE weight maps $[W_1, W_2, W_3, W_4]$. It can be found that different experts are assigned to focus on different areas.

5.3 Result Analysis

5.3.1 Quantitative Results. We summarize the quantitative results of different methods on the Kvasir-SEG and BKAI-IGH datasets in Table 2 and Table 3, and the observations are as follows. Our MoE-Polyp achieves good results and is superior to all competing methods in terms of all four evaluation metrics. For example, in Kvasir-Standard and BKAI-Standard, our MoE-Polyp outperforms PraNet in terms of mDice by 0.93% and 2.65%, respectively. While many models perform well in segmenting regular polyps, their effectiveness often diminishes when segmenting small polyps. Our MoE-Polyp not only excels at segmenting regular polyps but also shows significant improvements in segmenting small polyps compared to other models.

5.3.2 Qualitative Results. Figure 4 presents visual comparisons between our proposed MoE-Polyp and other competing methods across various challenging scenarios, including small size (row 1 and row 2), ambiguous boundaries (row 3 and row 4), and complex background (row 5 and row 6). As observed, MoE-Polyp effectively handles all these cases, achieving superior segmentation results.

5.4 MoE Visualization

To better understand the MoE decoder, in Figure 5, we visualize the MoE weight maps $[W_1, W_2, W_3, W_4]$ for the expert features in polyp segmentation. It is evident that the MoE design allows different weight maps to focus on distinct areas: W_1 emphasizes boundary information, W_2 concentrates on spatial details about the exterior polyp, W_3 contributes most and targets the main polyp body, and W_4 severs as an auxiliary to analyze the global background area.

6 Conclusion

In this paper, we propose MoE-Polyp, a novel framework for polyp image segmentation that employs a gating network to adaptively select and integrate suitable expert features, yielding precise segmentation results for polyps of varying sizes. Furthermore, we introduce a new benchmark specifically designed to enhance performance evaluation on small polyps. Extensive experiments demonstrate that the proposed MoE-Polyp framework excels at specializing in different features and accurately segmenting polyps.

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. 2009. Frequency-tuned salient region detection. In *CVPR*. IEEE, 1597–1604.

[2] Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. 2024. Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*. 801–818.

[4] Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. 2019. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *ICCV*. 3435–3444.

[5] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. 2017. Structure-measure: A new way to evaluate foreground maps. In *ICCV*. 4548–4557.

[6] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. 2020. Pranet: Parallel reverse attention network for polyp segmentation. In *MICCAI*. Springer, 263–273.

[7] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. 2019. Res2net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 2 (2019), 652–662.

[8] Along He, Kai Wang, Tao Li, Chengkun Du, Shuang Xia, and Huazhu Fu. 2023. H2Former: An efficient hierarchical hybrid transformer for medical image segmentation. *IEEE Transactions on Medical Imaging* (2023).

[9] Tianyu Hu, Yong Qi, Huafeng Wang, Yanqing Wang, Longzhen Wang, Minghua Du, and Xinyu Xiong. 2024. BCFNET: Boundary-Guided Semantic Cross Fusion for Polyp Segmentation. In *ISBI*. IEEE, 1–4.

[10] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. 2020. Kvasir-seg: A segmented polyp dataset. In *MMM*. Springer, 451–462.

[11] Wenxue Li, Wei Lu, Jinghui Chu, and Fugui Fan. 2023. LACINet: A lesion-aware contextual interaction network for polyp segmentation. *IEEE Transactions on Instrumentation and Measurement* (2023).

[12] Wenxue Li, Xinyu Xiong, Siying Li, and Fugui Fan. 2023. Hybridvps: Hybrid-supervised video polyp segmentation under low-cost labels. *IEEE Signal Processing Letters* (2023).

[13] Hanwei Liu, Rudong An, Zhimeng Zhang, Bowen Ma, Wei Zhang, Yan Song, Yujing Hu, Wei Chen, and Yu Ding. 2024. Norface: Improving Facial Expression Analysis by Identity Normalization. *arXiv preprint arXiv:2407.15617* (2024).

[14] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved Baselines with Visual Instruction Tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

[15] Liangliang Liu, Jianhong Cheng, Quan Quan, Fang-Xiang Wu, Yu-Ping Wang, and Jianxin Wang. 2020. A survey on U-shaped networks in medical image segmentations. *Neurocomputing* 409 (2020), 244–258.

[16] Zihan Liu, Hanyi Wang, Yaoyu Kang, and Shilin Wang. 2024. Mixture of Low-rank Experts for Transferable AI-Generated Image Detection. *arXiv preprint arXiv:2404.04883* (2024).

[17] Ange Lou, Shuyue Guan, Hanseok Ko, and Murray H Loew. 2022. CaraNet: context axial reverse attention network for segmentation of small medical objects. In *Medical Imaging 2022: Image Processing*, Vol. 12032. SPIE, 81–92.

[18] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. 2024. Segment anything in medical images. *Nature Communications* 15, 1 (2024), 654.

[19] Phan Ngoc Lan, Nguyen Sy An, Dao Viet Hang, Dao Van Long, Tran Quang Trung, Nguyen Thi Thuy, and Dinh Viet Sang. 2021. Neounet: Towards accurate colon polyp segmentation and neoplasm detection. In *ISVC*. Springer, 15–28.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*. Springer, 234–241.

[21] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *ICLR*.

[22] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. 2021. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* 71, 3 (2021), 209–249.

[23] Feilong Tang, Zhongxing Xu, Qiming Huang, Jinfeng Wang, Xianxu Hou, Jionglong Su, and Jingxin Liu. 2023. DuAT: Dual-aggregation transformer network for medical image segmentation. In *PRCV*. Springer, 343–356.

[24] Jun Wei, Yiwen Hu, Ruimao Zhang, Zhen Li, S Kevin Zhou, and Shuguang Cui. 2021. Shallow attention network for polyp segmentation. In *MICCAI*. Springer, 699–708.

[25] Jun Wei, Shuhui Wang, and Qingming Huang. 2020. F^3Net: fusion, feedback and focus for salient object detection. In *AAAI*, Vol. 34. 12321–12328.

[26] Xinyu Xiong, Zihuang Wu, Shuangyi Tan, Wenxue Li, Feilong Tang, Ying Chen, Siying Li, Jie Ma, and Guanbin Li. 2024. SAM2-UNet: Segment Anything 2 Makes Strong Encoder for Natural and Medical Image Segmentation. *arXiv preprint arXiv:2408.08870* (2024).

[27] Zhongxing Xu, Feilong Tang, Zhe Chen, Zheng Zhou, Weishan Wu, Yuyao Yang, Yu Liang, Jiyu Jiang, Xuyue Cai, and Jionglong Su. 2024. Polyp-Mamba: Polyp Segmentation with Visual Mamba. In *MICCAI*. Springer, 510–521.

[28] Yuqi Yang, Peng-Tao Jiang, Qibin Hou, Hao Zhang, Jinwei Chen, and Bo Li. 2024. Multi-Task Dense Prediction via Mixture of Low-Rank Experts. In *CVPR*.

[29] Guanghui Yue, Wanwan Han, Bin Jiang, Tianwei Zhou, Runmin Cong, and Tianfu Wang. 2022. Boundary constraint network with cross layer feature integration for polyp segmentation. *IEEE Journal of Biomedical and Health Informatics* 26, 8 (2022), 4090–4099.

[30] Guanghui Yue, Siying Li, Runmin Cong, Tianwei Zhou, Baiying Lei, and Tianfu Wang. 2023. Attention-guided pyramid context network for polyp segmentation in colonoscopy images. *IEEE Transactions on Instrumentation and Measurement* 72 (2023), 1–13.

[31] Ruifei Zhang, Peiwen Lai, Xiang Wan, De-Jun Fan, Feng Gao, Xiao-Jian Wu, and Guanbin Li. 2022. Lesion-aware dynamic kernel for polyp segmentation. In *MICCAI*. Springer, 99–109.

[32] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid scene parsing network. In *CVPR*. 2881–2890.

[33] Yijie Zhong, Bo Li, Lv Tang, Senyun Kuang, Shuang Wu, and Shouhong Ding. 2022. Detecting camouflaged object in frequency domain. In *CVPR*. 4504–4513.

[34] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. 2019. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging* 39, 6 (2019), 1856–1867.

[35] Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. 2024. MoVA: Adapting Mixture of Vision Experts to Multimodal Context. *arXiv preprint arXiv:2404.13046* (2024).