Lightweight Relational Embedding in Task-Interpolated Few-Shot Networks for Enhanced Gastrointestinal Disease Classification

1st Xinliu Zhong
Department of Biomedical Engineering
National University of Singapore
Singapore
xinliuzhong@u.nus.edu

2nd Leo Hwa Liang*
Department of Biomedical Engineering
National University of Singapore
Singapore
bielhl@nus.edu.sg

3rd Angela S. Koh
National Heart Centre Singapore;
Duke-NUS Medical School
Singapore
angela.koh.s.m@singhealth.com.sg

4th Yeo Si Yong*
Lee Kong Chian School of Medicine
Nanyang Technological University
Singapore
siyong.yeo@ntu.edu.sg

*Corresponding authors

Abstract*—Traditional diagnostic methods like colonoscopy are invasive yet critical tools necessary for accurately diagnosing colorectal cancer (CRC). Detection of CRC at early stages is crucial for increasing patient survival rates. However, colonoscopy is dependent on obtaining adequate and high-quality endoscopic images. Prolonged invasive procedures are inherently risky for patients, while suboptimal or insufficient images hamper diagnostic accuracy. These images, typically derived from video frames, often exhibit similar patterns, posing challenges in discrimination. To overcome these challenges, we propose a novel Deep Learning network built on a Few-Shot Learning architecture, which includes a tailored feature extractor, task interpolation, relational embedding, and a bi-level routing attention mechanism. The Few-Shot Learning paradigm enables our model to rapidly adapt to unseen fine-grained endoscopic image patterns, and the task interpolation augments the insufficient images artificially from varied instrument viewpoints. Our relational embedding approach discerns critical intra-image features and captures inter-image transitions between consecutive endoscopic frames, overcoming the limitations of Convolutional Neural Networks (CNNs). The integration of a light-weight attention mechanism ensures a concentrated analysis of pertinent image regions. By training on diverse datasets, the model's generalizability and robustness are notably improved for handling endoscopic images. Evaluated on Kvasir dataset, our model demonstrated superior performance, achieving an accuracy of 90.1%, precision of 0.845, recall of 0.942, and an F1 score of 0.891. This surpasses current state-of-the-art methods, presenting a promising solution to the challenges of invasive colonoscopy by optimizing CRC detection through advanced image analysis.

***Index Terms*—AI, Deep Learning, Few-Shot Learning**

I. Introduction

Gastrointestinal (GI) diseases encompass a range of conditions that can significantly impact health. Among these, colorectal cancer (CRC) is a notable concern due to its high prevalence and mortality rate when diagnosed late. Early detection is key to improving prognosis, with the potential to prevent up to 90% of deaths if identified promptly.

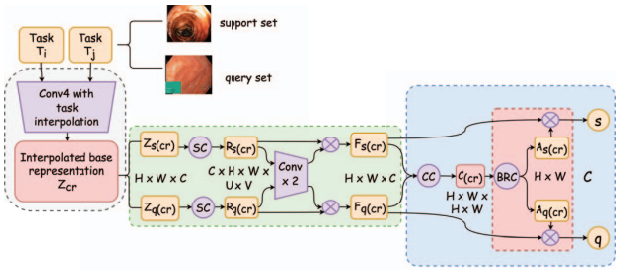

Fig. 1. Overall architecture of our proposed model. Given tasks, T, including support and query set images are sampled as input into the Conv4 encoder. Within the encoder, task interpolation is performed to enhance the diversity of the task set at a randomly chosen layer. Then the extracted support and query features, $Z_{s(cr)}$ and $Z_{q(cr)}$, are updated by self relational module separately. These features, $F_{s(cr)}$ and $F_{q(cr)}$, are integrated into the cross-relational module, which employs a lightweight bi-level routing attention mechanism to identify common areas of focus, $A_{s(cr)}$ and $A_{q(cr)}$, and derive the final outputs, s and q.

Endoscopic images are pivotal for diagnosing GI diseases, which predominantly arise from polyps in the lower digestive tract. The endoscope, a long tube with a fiber-optic camera, is used in this diagnostic procedure [1]. The surface morphological pattern of GI tract offers crucial clinical insights for surgical decisions and indicates disease aggressiveness.Admittedly, colonoscopy has several strengths, but it's not without its limitations. The intricate and varied nature of the colon and polyps, combined with the static size of the structuring element in the morphological operator, complicates segmentation, especially when vessels around the polyps evolve along the periphery of the liver [2], leading to substantial background variation, fuzzy frames and noise. Moreover, a multicenter prospective

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00157

study [3] has reported that up to 40% of CRC cases with deep submucosal invasion were misdiagnosed as superficial invasive cancer, underscoring the influence of the complexity of assessment theory and the subjectivity among endoscopists in impeding the accurate diagnosis of GI diseases.

Traditional computer vision techniques including Harris Corner Detection [4], SURF [5], and ORB [6], are heavily dependent on domain-specific knowledge and often fall short in capturing intrinsic image features in gastroenterology. Consequently, there's increasing interest in artificial intelligence (AI)-based techniques for more precise and objective GI disease identification. Deep Learning (DL) has led to a shift towards Convolutional Neural Networks [7] (CNNs) like Conv4 and ResNet12, which automatically extract meaningful image features hierarchically, outperforming traditional methods in classification tasks.

CNNs, foundational to various DL models, extract features through convolutions, influencing methods like LSTM, U-Net, and Inception. Dutta et al. [8] utilized a Tiny Darknet model, for efficient lesion detection. Another study [9] employed Xception [10], ResNet [11], and DenseNet [12] for ulcerative proctitis detection. Luo et al. [13] combined CNNs and Recurrent Neural Networks (RNNs) to diagnose ulcerative colitis from endoscopic images, enhancing accuracy with a spatial attention module. In endoscopic medical imaging, the challenges are multifaceted, ranging from data constraints to inherent limitations of conventional DLs. Addressing this, our novel DL model is tailored for endoscopic images, with the following key attributes:

- Traditional DL models, designed for generic images, falter with the specific textures and patterns of endoscopic images, especially when data is scarce in GI sector. By adopting a **few-shot learning** (FSL) paradigm, our model can **quickly understand and adapt** to the **unique characteristics** of endoscopic images, even with limited training data.
- Given the variability in endoscopic views due to different insertion angles and organ structures, over-fitting is a concern. Our model adds a **task interpolation** module in FSL to artificially diversify its training on various endoscopic perspectives, ensuring **robustness across different** endoscopic scenarios, overcoming the over-fitting issue.
- Endoscopic procedures often capture a series of images. Our model does not just process each image in isolation; it **understands the relationships and transition**s between consecutive endoscopic frames, offering a more comprehensive analysis.
- In endoscopy, minute details can be critical for diagnosis. While traditional CNNs might miss out on such details, our model, with its **advanced light-weight bi-level routing attention mechanism**, zeroes in on regions of interest (ROIs) in endoscopic images, ensuring no detail is overlooked and eliminating unwanted pixels.

II. Method

In this article, we introduce a DL approach for endoscopic data that can solve specific problems like fine-grained image patterns, small dataset size, as well as view variations. The DL methodology includes a feature extractor, FSL, data augmentation, relational embedding, and bi-level routing attention.

Fig.1 delineates the architectural overview of our proposed model, encapsulating the methodologies detailed subsequently — The rest of the section will be organized as follows: In Sec.II-A, the rationale behind opting for Conv4 over ResNet12 for strategic feature extraction is elucidated. The ensuing section, Sec.II-B, articulates the incorporation of the Few-Shot Learning (FSL) paradigm for handling the limited-size endoscopic image data. Sec.II-C explicates our approach to data augmentation through task interpolation, devised to synthesize multiple viewpoints. Moving forward, Sec.II-D describes the employment of both self- and cross-correlational embedding to discern relational representations within the CRC dataset, focusing on support and query images. Lastly, Sec.II-E details the deployment of a lightweight bi-level routing attention mechanism, aimed at efficiently attending to common ROIs within the data.

A. Feature Extractor

To efficiently process the complexities of endoscopic images, we chose Conv4 as our main feature extractor due to its balance of simplicity and performance. Although Conv4 and ResNet12 are both viable CNN architectures, Conv4's effectiveness in handling the detailed, resource-intensive nature of endoscopic image analysis made it our preferred choice.

Conv4, a lightweight and computationally efficient CNN architecture, is particularly advantageous for FSL tasks, especially in scenarios with limited computational resources. Comprising four convolutional blocks, each with a convolutional layer, batch normalization layer, and ReLU activation layer, Conv4 effectively learns and extracts meaningful features from images through a hierarchical approach. Its simplicity and efficiency make it a suitable encoder for our FSL model in endoscopic image analysis, where data is often scarce and computational efficiency is crucial.

B. Few-shot Learning

FSL is a DL approach that is particularly advantageous for training datasets with scarce information, a common challenge in CRC image classification within the domain of endoscopy. We used this paradigm of training and testing in building our model, as shown in Fig.2. This approach swiftly learns new concepts from minimal data, providing a viable solution to the data scarcity issue prevalent in medical imaging, especially in endoscopy where procurement of labeled data can be particularly challenging in less-than-ideal imaging conditions.

FSL builds a model by splitting the dataset D into D_{train} and D_{test}, ensuring the classes of the testing set should be unseen from the training set ($C_{train} \cap C_{test} = \varnothing$). Data from both sets are presented as tasks, each with a query set and a support set. For a task T_i, the support set $S_i = \{X_s^j, y_s^j\}_{j=1}^{NK}$

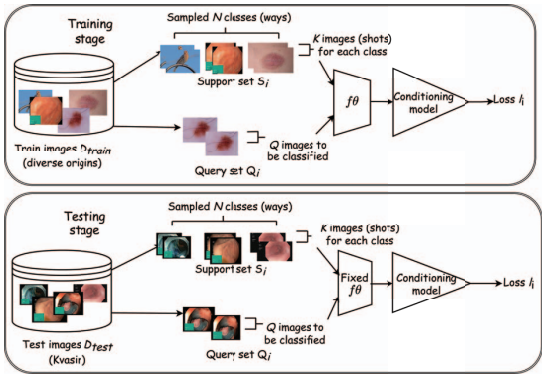

Fig. 2. Adopted FSL paradigm for our approach.

includes K images from each of N classes, and the query set $Q_i = \{X_q^j, y_q^j\}_{j=1}^{Q}$ comprises Q images, forming an N-way-K-shot task.

C. Data augmentation via task interpolation

Data augmentation, pivotal in enhancing ML models, especially in endoscopic image analysis, addresses limited data availability. For FSL in endoscopy, task interpolation is employed to generate additional domains, thereby augmenting the training data. This technique seeks to learn invariant representations and enhance training consistency by densifying the task distribution in a cross-task manner.

In a model f with L layers, the hidden representation of samples X at the l-th layer is $H_l = f_{\theta_l}(X)$, with $H_0 = X$ and L_s being the number of layers shared across all tasks. While meta-learning methods like MAML [14] share only a subset of the layers (i.e., $L_s < L$), metric-based methods like ProtoNet [15] share all layers (i.e., $L_s = L$). For a pair of tasks with their support and query sets sampled from the same label space, a layer l is randomly selected and task interpolation is applied separately on the hidden representations and corresponding labels of the support and query sets. The interpolated hidden representations and labels are $\tilde{H}_{s,l}^{cr} = \lambda H_{s,l}^{i} + (1-\lambda) H_{s,l}^{j}$, with $\lambda \in [0,1]$ sampled from a Beta distribution $Beta(\alpha, \beta)$ and subscript 'cr' indicating 'cross'. Both the hidden representations of support and query set are replaced by the interpolated ones in our task interpolation. The resulting enlarged task set is utilized as input for our FSL model, enhancing its capability to discern and classify endoscopic images across changing endoscope viewpoints.

D. Relational Embedding

Navigating through the intricacies of endoscopic images, especially within an FSL framework, demands a strategic approach to decipher and comprehend the subtle yet critical similarities and correlations among variables. This is particularly necessary in our target dataset, where the fine-grained medical images may appear superficially identical, and the data available is notably sparse. The challenge, therefore, is to unearth a relational structure that can simultaneously generalize across all embeddings, pinpointing ROIs that are crucial for precise classification and analysis.

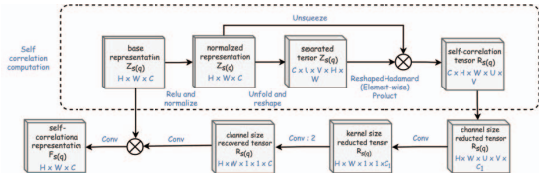

Fig. 3. Structure of self-correlational representation module. Base representation $Z_{s(q)}$ undergoes self-correlation computation by computing its Hadamard product $R_{s(q)}$. After a series of convolutions, the self-correlational representation $F_{s(q)}$ is derived to attend to pertinent areas inside the image.

The self-correlational representation (SCR) module [16], depicted in Fig.3, transforms the base representation Z from our feature extractor to highlight ROIs in a single image, thereby preparing a reliable input to the cross-correlational attention (CCA) module for analyzing feature correlations between a pair of different images. Inside SCR, we calculate a Hadamard product R within a neighbor window in a channel-wise manner as the self-correlation, followed by a series of convolutions, to produce the ultimate self-correlational feature. This feature, the same size as Z, is added back for reinforcement.

The CCA module takes a query-support pair from SCR and produces corresponding attention maps A_q and A_s, converting each representation to an embedding vector. It first computes the cross-correlation 4-dimensional tensor C by transforming both query and support representations, F_q and F_s, into more compact representations. Given potential unreliable correlations in C due to appearance variations in FSL, a convolutional matching process refines C into C', from which co-attention maps A_q and A_s are derived, highlighting relevant content between query and support.

E. Bi-level Routing Co-attention

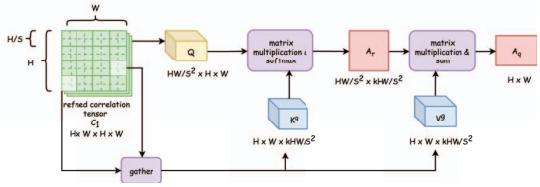

Fig. 4. Structure of bi-level routing co-attention module for A_q. The module simplifies attention calculations by gathering $topK$ key-value pairs K^g and V^g from coarse region-region affinity graph of the adjacency matrix A^r. Co-attention map for support set A_s is similar.

The attention mechanism, initially from Nature Language Processing (NLP), shows its ability as an alternative to convolutions in Transformers [17]. This structure can deal with global features (long-ranged dependencies), which is necessary for endoscopic images of GI. Inside the CCA module, we adopted co-attention mapping operations to reveal cross relations between the query and support.

In the standard Scaled Dot-product Attention calculation [17], inputs include a query matrix Q, and matrices for keys K and values V of dimension d_k. The output is computed

by applying a softmax function on the multiplication of the matrices as $Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$. Vision models often modify this by applying multi-head attention, concatenating parallel attention outputs for linear projection, with a complexity of $O(N^2)$ when there are N pairs of key-value pairs for each query in N queries.

To reduce the computational burden of attention, we implemented a bi-level routing attention mechanism [18], illustrated in Fig.4. This method selects the most relevant key-value pairs from a coarse region-region affinity graph in the adjacency matrix A^r, created by multiplying region-level queries Q^r and keys K^r. The routing index I^r is obtained using a top-k operation on A^r for pruning. This technique maintains co-attention capability while reducing computation costs by optimizing the region size.

III. Experiment

A. Datasets

TABLE I
Summary of Datasets Used in the Study

Dataset	*Images*	*Classes*	*Purpose*
Kvasir-v2 [19]	8000	8	GI disease classification
Hyper-Kvasir [20]	10,662	23	GI disease classification
ISIC 2018 [21][22]	10,208	8	Lesion classification
Cholec80 [23]	241,842	7	Surgery tool recognition
Mini-ImageNet [24]	60,000	100	General object classification

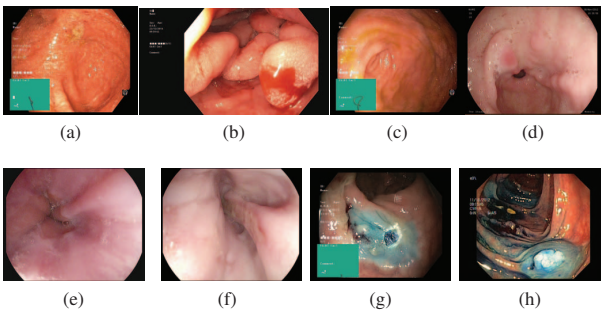

Fig. 5. Examples of Kvasir dataset [19]. (a)-(h): ulcerative colitis, polyps, normal cecum, normal pylorus, normal z-line, esophagitis, dyed resection margins, dyed lifted polyps.

To pretrain our model on endoscopic images, we utilized various datasets encompassing GI images, a broad-spectrum normal dataset, surgical tool images, and a skin disease dataset, enhancing the model's cross-domain generalization and reducing overfitting risk. The model underwent pretraining on ISIC 2018 [21][22], Cholec80 [23], and Mini-ImageNet datasets, followed by fine-tuning on the Hyper-Kvasir [20] dataset. This comprehensive fine-tuning involved 10,662 images across 23 GI disease classes, following the criteria in [1]. Incorporating a new full-connected layer head, the pretrained model was tailored for extracting endoscopic features. For testing, we employed the Kvasir-v2 dataset [19], a specialized multi-class image dataset for computer-aided GI disease detection, featuring 8 classes with 1000 images each, shown in fig.5. The datasets are summarized in Table.I.

B. Implementation details

Based on the configuration outlined in [25], we have eliminated the validation process. Additionally, we have adjusted our feature extractor, Conv4, by increasing the layer channels from 64 to 640, specifically to accommodate the relational embedding model structure. The experiments are performed under the N-way K-shot setting, where N = 2 for ISIC and N = 5 for the rest datasets, while K = 1 and number of query images is 15. The rationale Table.II of the experiment is listed below.

TABLE II
The Parameters of Experiment

Type	***Parameter***	***Value***
Feature Extractor	Conv4 layer	640
Adjuster	Optimizer	ADAM
	Learning rate	0.0001(fixed)
	Weight decay	0.002
Data loader	Batch size	128
	Saving episodes	500
	Total episodes	5000
FSL setting	Training way	5
	Testing way	2
	Shot	1

In data pre-processing, the images are resized to 92×92, randomly cropped to 84×84 and horizontally flipped for augmentation, and then its pixels normalized with mean values of $[125.3, 123.0, 113.9] \setminus 255.0$ and standard deviation values of $[63.0, 62.1, 66.7] \setminus 255.0$, inherited from ImageNet [24].

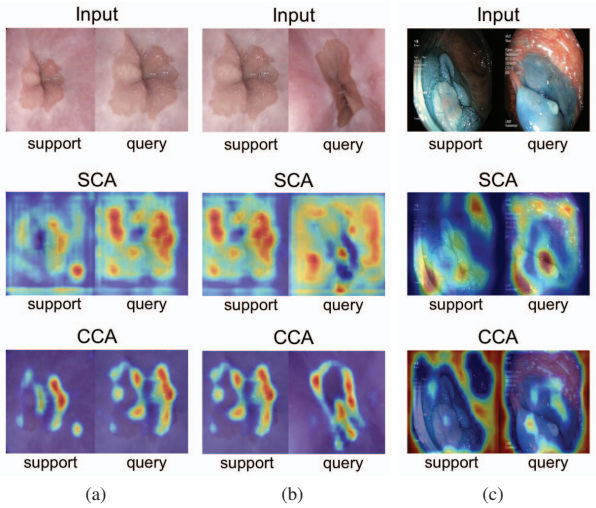

Fig. 6. Attention Heatmaps of SCA and CCA Modules with different source of support (left) and query (right) images. (a) Standard and zoomed views of a normal Z-line. (b) Different images of the same normal Z-line classification. (c) Different images of the same dyed lifted polyps classification.

After the images are preprocessed and grouped into batches of FSL tasks, they are augmented by cross-task interpolation.

Randomly select a layer l in the feature extraction process, and representations of the original query and support sets are replaced by the interpolated ones. These encodings are then input to the SCA module, which applies a channel-wise Hadamard product and convolutions to produce self-correlational representations F_q and F_s. The representations are processed by the CCA module into a 4-D tensor, from which we constructed affinity graphs for both query and support. The affinity graphs generated by our model are selectively pruned to identify the top-k routes, with resulting attentions illustrated in Fig.6.

Following the application of the two relational embedding modules, SCA and CCA, our model demonstrates a heightened capability to discern significant regions both within and across images. This proficiency is evident when the model processes pairs of images that either represent zoomed views of the same subject or belong to the same category but depict different subjects. In each case, the model adeptly navigates the classification task.

C. Comparison to the state-of-the-art methods

We compared our model to some state-of-the-art as well as several relevant methods on the endoscopic images. Quantitative results are shown in Table.III with metrics such as Accuracy, Precision, Recall, and F1-score.

TABLE III
CLASSIFICATION PERFORMANCE ON OUR PROPOSED MODEL AND OTHER MODELS

Methods	**Metrics**			
	Accuracy	*Precision*	*Recall*	*F1*
MAML [14]	0.792	0.610	0.633	0.621
ProtoNet [15]	0.775	0.662	0.694	0.678
Transformer [17]	0.870	0.738	0.812	0.773
ResNet50 [11]	0.812	0.701	0.794	0.745
Ours	**0.901**	**0.845**	**0.942**	**0.891**

Our proposed model demonstrates commendable performance, achieving an accuracy of 90.1% and an F-score of 0.89 on the dataset. This performance slightly surpasses that of Transformer and is 11% better than ProtoNet using Conv4.

IV. Ablation

To investigate the core effect of our main tenets, we conducted extensive ablation studies by omitting or substituting them with existing relevant methods. The different combinations are shown in Table.IV, demonstrating the effectiveness of the components in our proposed model. Additionally, we explored the influence of different pretraining datasets on our model's performance.

Results reported in Table.IV reveals following observations:

- Omitting ISIC 2018, Cholec80, or Mini-ImageNet during pretraining lowers performance, underscoring the importance of diverse, domain-specific data for robust feature learning and generalization.

TABLE IV
RESULTS OF ABLATION EXPERIMENTS

Model	***Accuracy***	***Inference Time***
Ours	**0.901**	0.52ms
Pretrained without ISIC 2018	0.869	0.53ms
Pretrained without Cholec80	0.873	0.53ms
Pretrained without Mini-ImageNet	0.894	0.52ms
Using ResNet12 instead of Conv4	0.844	0.54ms
Keeping Con4 64 layer channels and change relational layer to 64	0.898	0.52ms
Using MixUp [26] instead of task interpolation	0.878	0.45ms
Without task interpolation	0.870	0.43ms
Using vanilla attention instead of bi-level routing attention	0.895	0.61ms
Without attention	0.863	0.42ms
Without self correlational representation	0.886	0.52ms
Without cross correlational representation	0.857	**0.41ms**

- Utilizing Conv4 as opposed to ResNet12 and adjusting the relational layer's channel size while maintaining Conv4's 64-layer channels subtly impacts accuracy (0.901 vs. 0.844 and 0.898, respectively). This underscores the significance of the model architecture and relational layer channel dimensions in effectively extracting and relating features from endoscopic images, while also ensuring computational efficiency.
- Utilizing task interpolation, as opposed to MixUp or no interpolation (accuracy dips to 0.878 and 0.870, respectively), accentuates its role in bolstering model robustness by simulating varied viewpoints, vital for diverse endoscopic image scenarios.
- The implementation of bi-level routing attention, as opposed to vanilla or no attention (accuracy of 0.895 and 0.863, respectively), demonstrates its proficiency in effectively concentrating on pertinent features in GI images, thereby enhancing classification accuracy.
- Excluding self- or cross-correlational representation results in accuracy reductions to 0.886 and 0.857, respectively, highlighting their importance in capturing intra- and inter-image relationships, which is crucial for discerning subtle variations in endoscopic images.

V. Conclusion

Our DL model excels in analyzing endoscopic images with limited data, using FSL to quickly identify underlying patterns. It combats overfitting through task interpolation, simulating varied camera perspectives for better generalization.

The model's core strength is its ability to capture intra-image details and inter-image connections via self- and cross-correlational embedding. We enhanced it with a bi-level routing attention mechanism, making it lightweight yet focused in

image analysis. Trained on diverse datasets including Hyperkvasir and Mini-ImageNet, it achieved outstanding results on the Kvasir dataset, with 90.1% accuracy, 0.845 precision, 0.942 recall, and an F1 score of 0.891. This approach marks a significant step forward in endoscopic image analysis.

VI. Acknowledgment

This project is supported by the Lee Kong Chian School of Medicine - Ministry of Education Start-Up Grant.

References

[1] K. Ramamurthy, T. T. George, Y. Shah, and P. Sasidhar, "A novel multi-feature fusion method for classification of gastrointestinal diseases using endoscopy images," *Diagnostics*, vol. 12, no. 10, p. 2316,

[2] A. Akilandeswari, D. Sungeetha, C. Joseph, *et al.*, "Automatic detection and segmentation of colorectal cancer with deep residual convolutional neural network," *Evidence-Based Complementary and Alternative Medicine*, vol. 2022, 2022.

[3] Z. Lu, Y. Xu, L. Yao, *et al.*, "Real-time automated diagnosis of colorectal cancer invasion depth using a deep learning model with multimodal data (with video)," *Gastrointestinal Endoscopy*, vol. 95, no. 6, pp. 1186–1194, 2022.

[4] C. Harris and M. Stephens, "A Combined Corner and Edge Detector," in *Procedings of the Alvey Vision Conference 1988*, Alvey Vision Club, 1988, pp. 23.1–23.6.

[5] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Lecture notes in computer science*, vol. 3951, pp. 404–417, 2006.

[6] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*, Ieee, 2011, pp. 2564–2571.

[7] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, 2015. arXiv: 1409.1556 [cs.CV].

[8] A. Dutta, R. K. Bhattacharjee, and F. A. Barbhuiya, "Efficient detection of lesions during endoscopy," in *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VIII*, Springer, 2021, pp. 315–322.

[9] F. Zeng, X. Li, X. Deng, L. Yao, and G. Lian, "An image classification model based on transfer learning for ulcerative proctitis," *Multimedia Systems*, pp. 1–10, 2021.

[10] F. Chollet, *Xception: Deep learning with depthwise separable convolutions*, 2017. arXiv: 1610 . 02357 [cs.CV].

[11] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, 2015. arXiv: 1512 . 03385 [cs.CV].

[12] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, *Densely connected convolutional networks*, 2018. arXiv: 1608.06993 [cs.CV].

[13] X. Luo, J. Zhang, Z. Li, and R. Yang, "Diagnosis of ulcerative colitis from endoscopic images based on deep learning," *Biomedical Signal Processing and Control*, vol. 73, p. 103 443, 2022.

[14] C. Finn, P. Abbeel, and S. Levine, *Model-agnostic meta-learning for fast adaptation of deep networks*, Jul. 18, 2017. arXiv: 1703.03400[cs].

[15] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

[16] D. Kang, H. Kwon, J. Min, and M. Cho, *Relational embedding for few-shot classification*, 2021. arXiv: 2108. 09666 [cs.CV].

[17] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention Is All You Need*, arXiv:1706.03762 [cs], Aug. 2023.

[18] L. Zhu, X. Wang, Z. Ke, W. Zhang, and R. Lau, *BiFormer: Vision Transformer with Bi-Level Routing Attention*, arXiv:2303.08810 [cs], Mar. 2023.

[19] K. Pogorelov, K. R. Randel, C. Griwodz, *et al.*, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," in *Proceedings of the 8th ACM on Multimedia Systems Conference*, ACM, 2017, pp. 164–169.

[20] H. Borgli, V. Thambawita, P. H. Smedsrud, *et al.*, "HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy," *Scientific Data*, vol. 7, no. 1, p. 283, Aug. 2020, Number: 1 Publisher: Nature Publishing Group.

[21] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.

[22] N. Codella, V. Rotemberg, P. Tschandl, *et al.*, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)," *arXiv preprint arXiv:1902.03368*, 2019.

[23] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. de Mathelin, and N. Padoy, *Endonet: A deep architecture for recognition tasks on laparoscopic videos*, 2016. arXiv: 1602.03012 [cs.CV].

[24] O. Russakovsky, J. Deng, H. Su, *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[25] H. Yao, L. Zhang, and C. Finn, *Meta-learning with fewer tasks through task interpolation*, Mar. 17, 2022. arXiv: 2106.02695[cs].

[26] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.