UIT-Saviors at MEDVQA-GI 2023: Improving Multimodal Learning with Image Enhancement for Gastrointestinal Visual Question Answering

Triet M. Thai[1,2], Anh T. Vo[1,2], Hao K. Tieu[1,2], Linh N.P. Bui[1,2] and Thien T.B. Nguyen[1,2,*]

[1]*Faculty of Information Science and Engineering, Unviversity of Information Technology, Ho Chi Minh City, Vietnam*
[2]*Vietnam National University, Ho Chi Minh City, Vietnam*

Abstract

In recent years, artificial intelligence has played an important role in medicine and disease diagnosis, with many applications to be mentioned, one of which is Medical Visual Question Answering (MedVQA). By combining computer vision and natural language processing, MedVQA systems can assist experts in extracting relevant information from medical image based on a given question and providing precise diagnostic answers. The ImageCLEFmed-MEDVQA-GI-2023 challenge carried out a visual question answering (VQA) task in the gastrointestinal domain, which includes gastroscopy and colonoscopy images. Our team approached Task 1 - Visual Question Answering of the challenge by proposing a multimodal learning method with image enhancement to improve the VQA performance on gastrointestinal images. The multimodal architecture is set up with a BERT encoder and different pre-trained vision models based on convolutional neural network (CNN) and Transformer architecture for features extraction from question and endoscopy image. The result of this study highlights the dominance of Transformer-based vision models over the CNNs and demonstrates the effectiveness of the image enhancement process, with six out of the eight vision models achieving better F1-Score. Our best method, which takes advantages of BERT+BEiT fusion and image enhancement, achieves up to 87.25% accuracy and 91.85% F1-Score on the development test set, while also producing good result on the private test set with accuracy of 82.01%.

Keywords

visual question answering, multimodal learning, BERT, pre-trained models, gastrointestinal imaging, colonoscopy analysis, medical image processing

1. Introduction

The digestive system is one of the most complex and essential systems in the human body, consisting of various organs such as the mouth, stomach, intestines, and rectum. From the process of digestion in the stomach to the absorption of nutrients in the small and large intestines, and finally the elimination of waste through the rectum, the entire process involves the interaction and coordination of each organ to ensure the supply of nutrients and energy to the body. Any issues that occur in any part of the digestive system can directly impact the entire

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
*Corresponding author.
19522397@gm.uit.edu.vn (T. M. Thai); 19521226@gm.uit.edu.vn (A. T. Vo); 19521480@gm.uit.edu.vn (H. K. Tieu); 20521527@gm.uit.edu.vn (L. N.P. Bui); thienntb@uit.edu.vn (T. T.B. Nguyen)

gastrointestinal tract, such as inflammation of the intestines, digestive cancers, and diseases of the stomach and colon, especially colorectal diseases, which remain a significant concern for the healthcare community. According to estimates from the American Cancer Society[1], colorectal cancer ranks as the third leading cause of cancer-related deaths for both men and women in the United States. The projected numbers for colorectal cancer cases in the year 2023 are 106,970 new cases of colon cancer and 46,050 new cases of rectal cancer, with an estimated 52,550 deaths. However, it is important to note that the mortality rate from colorectal cancer has decreased over the past decade due to advancements in scientific and technological research. Screening techniques allow for the detection of abnormalities in the colon and rectum to be removed before they develop into cancer.

Clinical imaging techniques such as X-rays, computed tomography (CT), or ultrasound are often not highly effective in diagnosing pathological conditions in the colon. Therefore, colonoscopy remains the primary technique used for detection, screening, and treatment of gastrointestinal diseases. This method involves using a flexible endoscope, which is inserted through the anus and advanced into the colon. The real-time images of the colon obtained from the endoscopic device are displayed on a monitor, allowing the physician to observe and evaluate any abnormalities in the intestinal tract, the condition of the mucosal lining, and other structures within the colon.

Colonoscopy is considered the gold-standard screening procedure for examining and treating colorectal diseases. The endoscopic images contain a wealth of important information about the patient's condition. However, the effectiveness of the colonoscopy process can vary depending on the skills of the performer and the complexity of the endoscopic image analysis, which requires specialized knowledge and manual interpretation [1]. To improve the performance of colonoscopy in accurately detecting and classifying lesions, decision support systems aided by artificial intelligence (AI) are being rapidly developed. Among them, Visual Question Answering (VQA) is one of the most prominent techniques. Combining computer vision and natural language processing, VQA assists in extracting information from images, identifying abnormalities, and providing accurate answers to specific diagnostic questions. By integrating information from images and questions, VQA enhances the accuracy of lesion detection and classification, improves communication between users and images, and helps guide appropriate treatment strategies.

To successfully deploy VQA in the healthcare domain, in addition to algorithmic integration, a sufficiently large and diverse training dataset is required. Our research team participated in the VQA task of the ImageCLEFmed Medical Visual Question Answering on Gastrointestinal Image (MEDVQA-GI) [2] competition at ImageCLEF2023[3]. The contribution of the paper focused on performing the VQA task with a new dataset from ImageCLEFmed MEDVQA-GI. Specifically, we employed a multimodal approach for the VQA task (Task 1), combining information from two primary data sources: endoscopic images and textual questions. To achieve a good performance on the VQA task with the provided dataset, we first performed an efficient image preprocessing steps, which involved specular highlights inpainting, noise, and black mask removal to enhance the image quality. Subsequently, we conducted experiments and compared the performance of various image feature extraction models based on CNN and Transformer using both raw

[1] https://www.cancer.org/cancer/types/colon-rectal-cancer/about/key-statistics.html.

and enhanced image data. The final results, with accuracy up to 87.25% on the development test set and 82.01% on the private test set, demonstrate the potential of the proposed method in improving the performance of VQA systems in the field of gastrointestinal endoscopy imaging in general and colonoscopy in particular.

2. Background and Related Works

2.1. Colonoscopy Image Analysis

With the advancement of modern advanced technology, AI has made significant contributions to the field of healthcare, specifically in the progress of the colonoscopy examination process. Currently, two potential approaches with AI being utilized for colonoscopy image analysis, including Computer-Aided Detection (CAD) and Deep Learning (DL) systems. In the CAD approach, the system utilizes image processing algorithms to improve the performance of endoscopic procedures, enabling physicians to easily detect lesions in hard-to-identify locations and reduce the chances of misdiagnosis [4]. On the other hand, the DL-based system employs a deep learning model trained on specific datasets, which enhances the accuracy of lesion detection compared to the CAD-based system [5]. However, developing algorithms for automatic analysis and anomaly detection in endoscopic images requires preliminary image preprocessing to address various factors, such as specular highlights, interlacing or artefacts that impact the system's performance [6].

2.2. Preprocessing Methods for Colonoscopy Images

In reality, the quality of endoscopy images depends on various factors such as the skill of the performing physician, limitations of the equipment, and certain environmental conditions. Some common difficulties in processing endoscopy images include black masks, ghost colors, interlacing, specular highlights, and uneven lighting [7]. Black masks are the occurrence of a black border around the edges of the image due to the use of lenses in the endoscopy system that have a black frame surrounding the edges. This frame can hinder the development of algorithms. To address this issue, techniques such as restoration, thresholding, cropping, or inpainting are necessary. Specular highlights, which are bright spots reflected from tumors or polyps captured by the camera, can disrupt the algorithms. Therefore, to remove them, we can employ detection or inpainting methods. Additionally, for issues like interlacing, ghost colors, and uneven lighting, segmentation methods can be applied to achieve optimal results [6] [8] [9]. Overall, preprocessing steps play a crucial role in mitigating the challenges commonly encountered with colonoscopy images. The mentioned techniques will help improve the overall quality of the images, thereby enhancing the performance of analysis and diagnosis.

2.3. Medical Visual Question Answering

Medical visual question answering (MedVQA) is an important field in medical AI that combines VQA challenges with healthcare applications. By integrating medical images and clinically relevant questions, MedVQA systems aim to provide plausible and convincing answers. While

VQA has been extensively studied in general domains, MedVQA presents unique opportunities for exploration. Currently, there are 8 publicly available MedVQA datasets, including VQA-MED-2018 [10], VQA-RAD [11], VQA-MED-2019 [12], RadVisDial [13], PathVQA [14], VQA-MED-2020 [15], SLAKE [16], and VQA-MED-2021 [15]. These datasets serve as valuable resources for advancing MedVQA research.

The basic framework of MedVQA systems typically contains an image encoder, a question encoder, a fusion algorithm, and an answering component. Other frameworks may exclude the question encoder when the question is simple. Common choices for image encoder are ResNet [17] and VGGNet [18] that are pre-trained on ImageNet dataset [19]. For language encoders, Transformer-based architectures such as BERT [20] or BioBERT [21] are commonly applied because of their proven advantages, besides the Recurrent Neural Networks (LSTM [22], Bi-LSTM [23], GRU [24]). The fusion stage, the core component of VQA methods, has typical fusion algorithms, including the attention mechanism and the pooling module. Common attention mechanisms are the Stacked Attention Networks (SAN) [25], the Bilinear Attention Networks (BAN) [26], or the Hierarchical Question-Image Co-Attention (HieCoAtt) [27]. Most multimodal pooling practices are concatenation, sum, and element-wise product. The attention mechanism can aggregate with the pooling module. The answering component has two modes of output depending on the properties of the answer. The classification mode is used if the answer is brief and limited to one or two words. Otherwise, if the response is in free-form format, the generation modules such as LSTM or GRU are taken into account. There are additional techniques to the basic concept, for instance, Sub-task strategy, Global Average Pooling [28], Embedding-based Topic Model, Question-Conditioned Reasoning, and Image Size Encoder.

3. Task and Dataset Descriptions

3.1. Task Descriptions

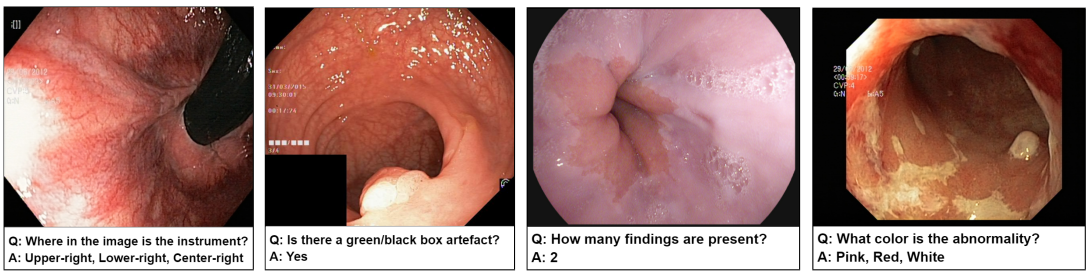
Identifying lesions in endoscopy images is currently one of the most popular applications of artificial intelligence in the medical field. For the task at ImageCLEFmed-MEDVQA-GI-2023 [2], the main focus will be on VQA and visual question generation (VQG). The main goal is to provide support to healthcare experts in diagnosis by combining image and text data for analysis. The task consists of three sub-tasks:

1. **VQA (Visual Question Answering):** For the visual question answering part, participants are required to generate a textual answer to a given textual question-image pair. This task involves combining endoscopy images from the dataset with textual answers to respond to questions.
2. **VQG (Visual Question Generation):** This is the reverse task of VQA, where participants need to generate textual questions based on given textual answers and image pairs.
3. **VLQA (Visual Location Question Answering):** Participants are provided with an image and a question, and they are required to provide an answer by providing a segmentation mask for the image.

In this study, our team only focuses on the VQA task (Task 1) for the provided endoscopy image dataset. In general, we receive a textual question along with the corresponding image,

Table 1
Questions and sample answers from ImageCLEFmed-MEDVQA-GI-2023 dataset

ID	Questions	Sample Answers
0	What type of procedure is the image taken from?	"Colonoscopy", "Gastroscopy"
1	Have all polyps been removed?	"Yes", "No", "Not relevant"
2	Is this finding easy to detect?	"Yes", "No", "Not relevant"
3	Is there a green/black box artifact?	"Yes", "No"
4	Is there text?	"Yes", "No"
5	What color is the abnormality?	"Red", "Pink", "White, Yellow", ...
6	What color is the anatomical landmark?	"Red", "Red, White", "Pink, Red, grey", ...
7	How many findings are present?	"0", "1", "2", "3", "4", "5", ...
8	How many polyps are in the image?	"0", "1", "2", "3", "4", "5", ...
9	How many instruments are in the image?	"0", "1", "2", "3"
10	Where in the image is the abnormality?	"Center", "Lower-left", "Lower-right, Center-right", ...
11	Where in the image is the instrument?	"Center", "Lower-left", "Lower-right, Center-right", ...
12	Are there any abnormalities in the image?	"No", "Polyp", "Ulcerative colitis", "Oesophagitis", ...
13	Are there any anatomical landmarks in the image?	"No", "Z-line" "Cecum", "Ileum", "Pylorus", "Not relevant"
14	Are there any instruments in the image?	"No", "Tube", "Biopsy forceps", "Metal clip", 'Polyp snare, Tube', ...
15	Where in the image is the anatomical landmark?	"Center", "Lower-left", "Lower-right, Center-right", ...
16	What is the size of the polyp?	"< 5mm", "5-10mm", "11-20mm", ">20mm", "Not relevant", ...
17	What type of polyp is present?	"Paris ip", "Paris iia", "Paris is", "Paris is, Paris iia", ...



(a) No abnormality (b) Polyp (c) Oesophagitis (d) Ulcerative Colitis

Figure 1: Illustrations of question-answer pairs along with common abnormalities in gastrointestinal image from ImageCLEFmed-MEDVQA-GI-2023 dataset

and the main task is to generate accurate and appropriate answers based on information from both sources. For example, for an image containing a colon polyp with the following question, "Where in the image is the polyp located?", the proposed VQA system should return answer giving a textual description of where in the image the polyp is located, like upper-left or in the center of the image.

3.2. Dataset Information

The new dataset released for the ImageCLEFmed-MEDVQA-GI-2023 challenge is based on the HyperKvasir dataset [29], the largest gastrointestinal collections with more than 100,000 images, with the additional question-and-answer ground truth developed by medical collaborators. The development set and test set include a total of 3949 images from different procedures such as gastroscopy and colonoscopy, spanning the entire gastrointestinal tract, from mouth to anus. Each image has a total of 18 questions about abnormalities, surgical instruments, normal

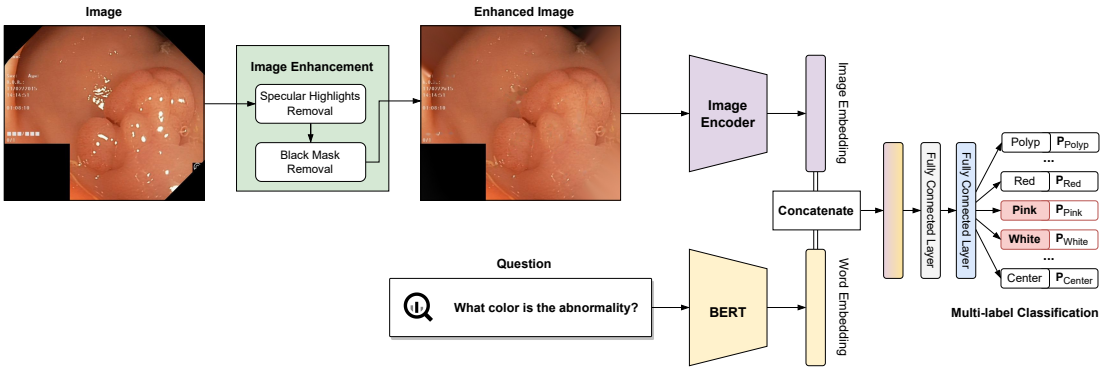


Figure 2: An overview of the multimodal architecture with image enhancement for VQA challenge

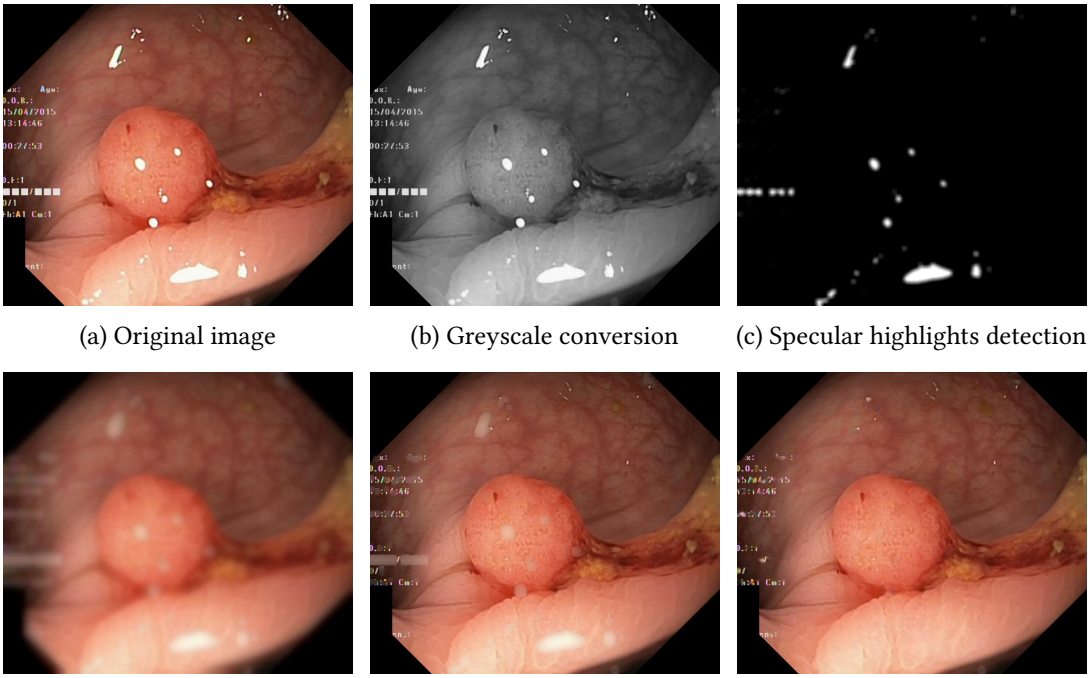
findings and other artefacts, with multiple answers possible for each, as shown in Table 1. Not all questions will be relevant to the provided image, and the VQA system should be able to handle cases where there is no correct answer. Figure 1 depicts several examples of question-answer pairs on common abnormalities in gastrointestinal tract, such as Colon Polyps, Oesophagitis, and Ulcerative Colitis. As shown in Figure 1d, there are three possible answers to the question "What color is the abnormality?": "Pink," "Red," and "White", and a typical VQA system should be able to identify all three colors. In general, the image may contains a variety of noise and components that locates across abnormalities, such as highlight spots or instruments, which pose a significant challenge in developing efficient VQA systems for gastrointestinal domain.

4. The Proposed Approach

The method used in this study is based on a standard framework that is commonly used to tackle general VQA problems. Figure 2 depicts an overview of the proposed method for ImageCLEFmed-MEDVQA-GI-2023 dataset. In general, the VQA architecture employs powerful pre-trained models to extract visual and textual features from image-question pairs, which are then combined into a joint embedding using a fusion algorithm and passed to a classifier module to generate the appropriate answer. To improve the quality of the region of interest and achieve better VQA performance, the original image is passed through a series of enhancement procedures before being fed into the image encoder for features extraction.

4.1. Image Enhancement

The purpose of the image pre-processing and enhancement steps is to remove noise and artifacts, which are frequently caused by the equipment used in diagnostic or environmental difficulties. Some of the major problems to be mentioned are black mask, specular highlights, interlacing or uneven lighting. The impact of these elements, such as black mask and specular highlights, is significant since they, like the polyp, create valley information and affect the performance of polyp localization, causing the VQA system to generate incorrect answers.



(a) Original image (b) Greyscale conversion (c) Specular highlights detection

(d) Average smoothing (e) Initial restoration (f) Inpainted using Telea algorithm

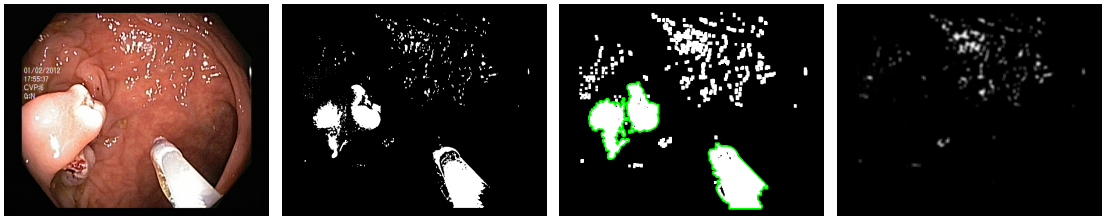
Figure 3: An overview of stages of the specular highlights inpainting method

This study employs pre-processing and enhancing methods to cope with specular highlights and black mask in colonoscopy image, which are prevalent artifacts in the dataset provided. The desired outcome is an enhanced image with no specular reflection or black frame while retaining the visual features of the region of interest.

4.1.1. Specular Highlights Removal

The removal of specular highlights from colonoscopy image includes two sequential processes: detection of specular highlights and highlights inpainting. Figure 3 depicts the overall procedure of the method, the outcome of which is generally based on the combination of Telea inpainting algorithm with initial image restoration after several modification steps.

Specular highlights detection First, it is necessary to convert the image from the original RGB channel to grey scale to process the subsequent procedure. Rather than adaptive thresholding, the proposed approach employs standard thresholding method with a fixed threshold value to identify specular highlights in all images. This is due to the gastrointestinal image's varied textures and components, and if not done properly, may result in information loss. Some samples of the dataset contain text, high exposure regions and brightly colored instrument, as described in Figure 4. Aside from text in white color, high exposure regions are parts of specular highlights that received excessively high intensity compared to regular highlight spots, while the instruments are sometimes in white or blue color. After thresholding, these factors



(a) Original image (b) Thresholding (c) Detection of undesired regions (d) Preprocessed mask

Figure 4: An illustration of specular highlights detection from a colonoscopy image that contains text, high exposure regions and a white instrument.

may emerge in the mask, as shown in Figure 4b, and affect the inpainting outcome. Thus, the following step is to remove these undesired elements from the mask in order to assure consistency. To cope with these problems, two directions are considered, either to perform segmentation for text, polyp and instrument, separately, or remove the parts that meet certain size threshold. For simplicity, the second approach is used in this study.

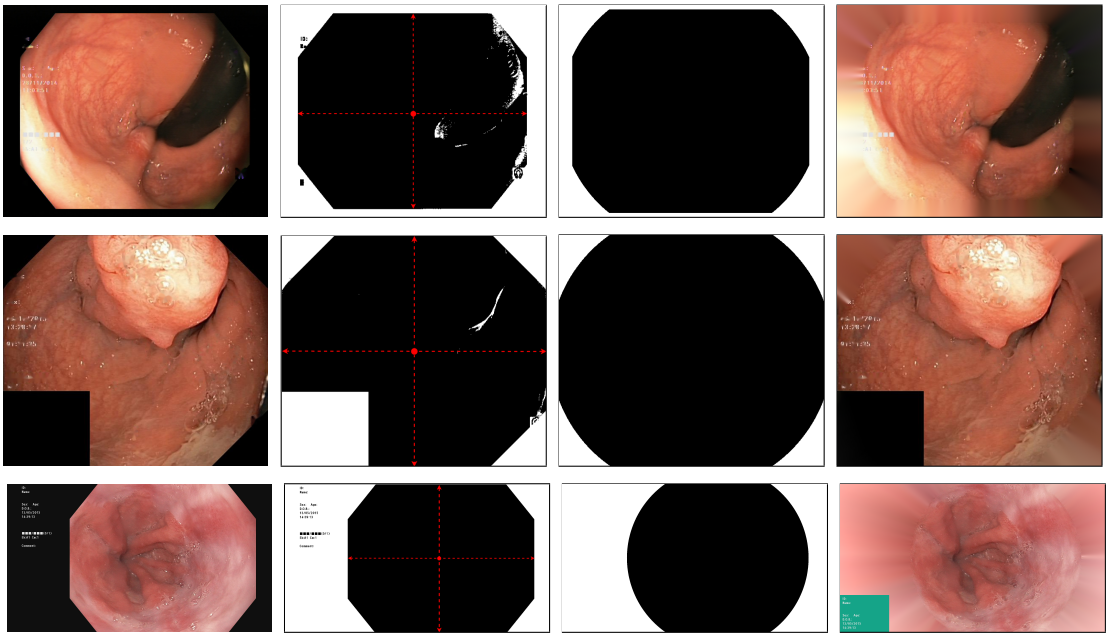
The preprocessing step consists of several morphology transformations interspersed by contour detection and removal. More specifically, a dilation operation with kernel size 3×3 is performed initially to connect the pixels related to undesirable parts. Among the obtained contours, those whose scaled area following the Modified Z-scores formula [30], as shown in Formula 1, exceeds 17.0 are removed from the mask. The mask is then passed into another erosion module with the same settings to restore the initial highlights intensity. Finally, Gaussian filter of size 19×19 is applied to reduce the intensity of highlights area and improve the inpainting performance.

$$S_i = \frac{|s_i - \tilde{s}|}{MAD} \tag{1}$$

where:

- S_i: is the scaled area of contour i based on modified Z-score.
- s_i: is the area of contour i
- $\tilde{s}$: is the median area of all contours
- $MAD = median(|s_i - \tilde{s}|), \forall i = 1..n$: is the Median Absolute Deviation of contour areas

Highlights inpainting Once the mask of specular highlights has been achieved, the image regions indicated by the mask are then reconstructed through an efficient inpainting operation. First, a filter of size 3×3 slides across every pixels of the original image and calculate the average value. The process is repeated N times to ensure a desirable outcome. We then perform an initial restoration on the image by directly replacing its pixels under the specular highlights mask with pixels from the blur image. Despite the drastically reduced intensity, specular highlight spots still remains in the reconstructed image, as shown in Figure 3e. To obtained the final result, Telea algorithm [31], a powerful image inpainting strategy, is applied to eliminate the remaining noisy and dim highlights. The inpainted image is noticeably higher in quality, with specular highlights removed without negatively impacting other areas of the image.



(a) Image with black mask (b) Border detection (c) Artificial mask (d) Enhanced Image

Figure 5: Stages of black mask removal process. The first row illustrates an image with a standard black mask, the second row depicts an image containing a black square and the last row contains image with black mask marked as black box artefact.

4.1.2. Black Mask Removal

Previous research has shown that black masks do generate valley information, which can reduce polyp localization performance. Based on this, we propose a black mask removal strategy for the VQA task that still retains black box information in order to answer the question "Is there a green/black box artefact?". In general, an artificial mask of black frame is initially created based on its border width, and then the inpainting operation is performed to remove the black frame from the image. The overall procedure is described in Figure 5. Our method does not use cropping or thresholding directly to detect and remove the black mask because it may contain the black box artifact, shadow regions, or black instrument, the removal of which causes information loss and decreases VQA performance.

To detect the border width, we first perform a grey scale conversion and inverse thresholding with erosion operation to remove noise, and then measure the distance from each edge of the image to the nearest pixel that does not belong to the mask. After determining the width of the border, the crucial step of the method is to create an artificial mask with internal octagon shape. This can be done by creating two sub-masks, one rectangle and one circle, followed by a bitwise OR operation to combine them into the final mask, as show in Figure 5c. The circle mask is created with a center point based on the information of border width and a radius calculated by multiplying the ordinate of the center point by a value σ ($\sigma > 1$). In some cases, the final mask is not always octagonal, as shown in the last example, but it still covers the main region of interest. Finally, the inpainting of black mask is completed using the same procedure as

described in the previous section for specular highlights, giving the final enhanced image with black mask removed. If a black box artefact exists in the bottom-left corner, as shown in the second example, it will not be significantly affected as long as its size is greater than the area of the mask at the respective position. For images containing an expanded black mask labeled as black box artefact, we process by creating a simulated green box that contains the text and placing it in bottom-left corner. By doing so, the text and box artefact information still remain after the inpainting procedure. Though the obtained results are quite satisfactory, there are still some cases where the mask is not completely removed and need further processing steps.

4.2. Multimodal Fusion Architecture

Since this study focus mainly on the VQA task, the architecture should be capable of extracting meaningful features from the question and corresponding image, and incorporating them to give the correct answer. Our multimodal fusion architecture is set up with important components such as an image encoder for feature extraction from images, a text encoder for features extraction from questions, a fusion algorithm for unifying modalities and a classifier for producing the appropriate answer. The proposed approach uses pre-trained Bidirectional Encoder Representations based on Transformers (BERT) [20] to extract textual features from questions. As a bidirectional model, it can learn the meaning of words in a sentence by considering both the words that come before and after them. With massive pre-training data, BERT can be fine-tuned and achieved state-of-the-art results on a number of natural language processing (NLP) benchmarks. For features extraction from the images, this study set up and experiment with eight different pre-trained models that are belong to two main concepts:

- CNN-based architectures including **Resnet152** [32], **Inception-v4** [33], **MobileNetV2** [34] and **EfficientNet** [35]. The group of models take advantage of traditional CNN's components such the convolutional layer, pooling layer, residual block and fully connected layer to achieve significant result in computer vision field. The training of CNN-based model is more efficient with less computational resources compared to new approaches based on Transformers.
- Transformer-based architectures including **ViT** [36], **DeiT** [37], **Swin Transformer** [38] and **BEiT** [39]. The family of models leverages a massive amount of training data and Transformer's multi-head self-attention for a game-changing breakthrough in the computer vision field. ViT (Vision Transformer) and other models inspired from it initially encodes the image as patch embeddings and pass them into a regular Transformer Encoder for feature extraction, which is similar to text data. Currently they are considered as the prominent architectures to achieve state-of-the-art performance on a variety of tasks in computer vision such as image classification, object detection, and semantic image segmentation.

After obtaining the embeddings of text and image, a multimodal fusion method based on concatenation is used to combine these features along the embedding dimension. The unified embedding matrix is then passed through an intermediate fully connected layer with drop out 0.5 and ReLU activation followed by a classification layer to produce the final output. Because there can be more than one appropriate answer for each question, we approach the VQA task as

Table 2
Statistics of multimodal fusion with pre-trained vision and language models for the VQA challenge

Models	Version	Spaces vision model name	# Parameters
BERT+ViT	base	"google/vit-base-patch16-224-in21k"	196M
BERT+DEiT	base	"facebook/deit-base-distilled-patch16-224"	196M
BERT+Swin	base	"microsoft/swin-base-patch4-window7-224-in22k"	197M
BERT+ BEiT	base	"microsoft/beit-base-patch16-224-pt22k-ft22k"	196M
BERT+ResNet152	v1.5	"microsoft/resnet-152"	169M
BERT+Inception	v4	"inception_v4"	153M
BERT+MobileNet	V2	"google/mobilenet_v2_1.0_224"	112M
BERT+EfficientNet	b3	"google/efficientnet-b3"	121M

a multi-label classification problem. To successfully train the proposed arhitecture, multi-label binarization is used to encode a list of all possible answers into a binary vector. Furthermore, the final layer is configured with sigmoid activation function to return an output vector of the same size containing the corresponding probability for each class.

5. Experimental Setup

5.1. Data Preparation

The development set released for the VQA challenge contains 2000 images of gastroscopy and colonoscopy procedures. In order to experiment and evaluate our method, we randomly divided the provided development set into three parts: train, validation, and test, with 1600 images for training and 200 images for each validation and test set. The data preparation process is designed to ensure that each abnormality has the same proportion in the training, validation, and testing sets, and that each image contains all 18 questions. This produces 28,800 question-answer pairs on the training set, 3600 pairs for validation and 3600 pairs for test.

All images from development set and private test set are first passed into an image enhancement block, where numerous image preprocessing methods are applied to remove specular highlights and black mask from the images. The enhanced results are then used as input in the training and testing of the proposed VQA model.

5.2. Experiment Configurations

Many experiments are carried out in order to evaluate the performance of the proposed methods toward the ImageCLEFmed-MEDVQA-GI-2023 challenge. Specifically, each pre-trained vision model is initialized and experimented as an image encoder and unify with BERT encoder through concatenation fusion for multimodal learning. Table 2 gives the general information of pre-trained models used in this study including vision model name, version and number of parameters for each fusion model. Through experiments, we can discover the potential and limitation of each model for the VQA task and thus, choose the best method for giving the final prediction on the private test set of the competition.

To achieve a comparative result, we set up the same hyperparameters for all experiments. The models are trained in 15 epochs with batch size of 64. We utilize the Adam optimizer [40] using weighted decay with an initial learning rate of 5e-5 and a linear scheduler to decrease learning rate 6.67% after each epoch. Since we approach the VQA task as multi-label classification, the output layer is configured to return a tensor containing probabilities of answers, where the final predicted answers for each question can be achieved using threshold value of 0.5. Due to this, the BCEWithLogitsLoss function, which combines a Sigmoid layer and the BCELoss, is applied in the training process. After each epoch, the training loss and validation loss are calculated, and the performance are then evaluated on classification metrics such as accuracy, precision, recall and F1-Score. To ensure a meaningful result for multi-label classification, the metrics are calculated using ground truth and prediction sets of binary vectors, in which recall, precision and F1-scores should be calculated on each sample and find their average. The model's state that obtains best F1-Score is used for prediction in the testing phase.

The proposed architecture are implemented in PyTorch and trained on the Kaggle platform with hardware specifications: Intel(R) Xeon(R) CPU @ 2.00GHz; GPU Tesla P100 16 GB with CUDA 11.4.

6. Experimental Results

Table 3
Comparative performance of the multimodal fusion method with vision models on the development test set.

Vision Models	Accuracy	Precision	Recall	F1-Score
No image enhancement				
ResNet152	0.8419	0.8917	0.8867	0.8857
Inception-v4	0.8619	0.9133	0.9067	0.9067
MobileNetV2	0.8444	0.8932	0.8951	0.8906
EfficientNet-B3	0.8581	0.9065	0.9049	0.9023
ViT-B/16	0.8636	0.9134	0.9089	0.9078
DeiT-B	0.8611	0.9100	0.9026	0.9033
Swin-B	**0.8664**	0.9152	**0.9094**	**0.9090**
BEiT-B	0.8647	**0.9158**	0.9068	0.9074
With image enhancement				
ResNet152	0.8453 ↑	0.8942	0.8894	0.8885 ↑
Inception-v4	0.8625 ↑	0.9121	0.9073	0.9071 ↑
MobileNetV2	0.8422 ↓	0.8935	0.8882	0.8867 ↓
EfficientNet-B3	0.8572 ↓	0.9081	0.9079	0.9046 ↑
ViT-B/16	0.8631 ↓	0.9126	0.9086	0.9073 ↓
DeiT-B	0.8625 ↑	0.9122	0.9052	0.9055 ↑
Swin-B	0.8717 ↑	0.9245	0.9159	0.9168 ↑
BEiT-B	**0.8725** ↑	**0.9253**	**0.9184**	**0.9185** ↑

Table 4
Performance evaluation of BERT+BEiT fusion with image enhancement for each question on the development test set and private test set

Question	Development Test Set				Private Test Set
	Accuracy	Precision	Recall	F1-Score	Accuracy
Are there any abnormalities in the image?	0.9700	0.9750	0.9725	0.9733	0.8091
Are there any anatomical landmarks in the image?	0.9300	0.9300	0.9300	0.9300	0.6940
Are there any instruments in the image?	0.9050	0.9275	0.9200	0.9217	0.7688
Have all polyps been removed?	0.9550	0.9575	0.9600	0.9583	0.9721
How many findings are present?	0.8400	0.8400	0.8400	0.8400	0.7807
How many instrumnets are in the image?	0.9650	0.9650	0.9650	0.9650	0.8901
How many polyps are in the image?	0.9650	0.9650	0.9650	0.9650	0.9577
Is there a green/black box artefact?	0.9500	0.9500	0.9500	0.9500	0.9732
Is there text?	0.9250	0.9250	0.9250	0.9250	0.8787
Is this finding easy to detect?	0.8900	0.8900	0.8900	0.8900	0.8044
What color is the abnormality?	0.5800	0.9025	0.8563	0.8597	0.4969
What color is the anatomical landmark?	0.9400	0.9400	0.9400	0.9400	1.0000
What is the size of the polyp?	0.8600	0.8650	0.8700	0.8667	0.8535
What type of polyp is present?	0.8650	0.8800	0.8725	0.8750	0.8132
What type of procedure is the image taken from?	1.0000	1.0000	1.0000	1.0000	0.9938
Where in the image is the abnormality?	0.6600	0.9251	0.8805	0.8842	0.5872
Where in the image is the anatomical landmark?	0.7150	0.8847	0.8848	0.8766	0.7203
Where in the image is the instrument?	0.7900	0.9332	0.9096	0.9125	0.7688
All	**0.8725**	**0.9253**	**0.9184**	**0.9185**	**0.8201**

The comparative result of different pre-trained image model on the testing set is shown in Table 3. It is clear that, with no image enhancement, Swin-B achieves the best result with 86.64% accuracy and 90.90% F1-Score while BEiT-B gives a slightly lower performance with accuracy of 86.47% and 90.74% F1-Score. CNN-based vision models have acceptable results, but cannot be compared with the result of Transformer architecture models.

With image enhancement, six out of eight vision models from both CNN and Transformer architectures achieve a better performance on F1-Score metric. BEiT-B has an outstanding result with accuracy and F1-Score of 87.2% and 91.85%, respectively. Overall, the enhancement process helps to improve the F1-Score at least 0.4% and up to 1.11% on VQA performance. The result of the convolutional models is still under when compared with Transformers architecture models.

We found that the BERT and BEiT fusion (BERT+BEiT) with image enhancement is the best method of our approach and use it for prediction in final private test phase. Our method obtains a good result on the private test set with an accuracy of 82.01%. Table 4 illustrates the performance evaluation of BERT+BEiT fusion on each question from the development test set compared with the private test set. In general, there are 14/18 questions with predicted answers achieve greater than 80% accuracy on the development test set, while 11/18 questions on the private test set achieve the same result. Our method still struggles to produce full and precise answers for questions with multiple answers, such as "What color is the abnormality?" or questions that refer to the location of the abnormality, anatomical landmark, and instrument.

7. Conclusion and Future Works

Along with performing image enhancement, we also set up and experimented using various powerful pre-trained image models together with the BERT encoder for our proposed multimodal architecture in the VQA task at ImageCLEFmed-MEDVQA-GI-2023 [2]. The visual enhancement steps, which include specular hightlights and black mask removals, help improve multimodal learning performance on the dataset by up to 1.11% F1-Score. Our best method, BERT+BEiT fusion with image enhancement, achieved 87.25% on development test set and 82.01% on the private test set by the accuracy. Through performance analysis, there are question cases that require multiple positions or colors in the answer, which are our limitations in this study. In summary, there are factors that have significant impact on our solution for the VQA task such as answer imbalance, noise, and artifacts.

Our future research for this task is to improve the accuracy of the model in giving the correct answer by enriching the features from images and questions through instrument segmentation and polyp localization with methods such as U-net [41], ResUnet++ [42] developed on object-specific datasets such as Kvasir-Instrument [43] and Kvasir-seg [44]. Other advanced colonoscopy image preprocessing techniques such as interlacing removal or uneven lighting removal can be examined to improve the image quality. From the proposed system, an intelligent chatbot application can be implemented for question-answering from medical images and help improve colonoscopy analysis.

Acknowledgments

This research was supported by The VNUHCM University of Information Technology's Scientific Research Support Fund.

References

[1] D. K. Rex, P. S. Schoenfeld, J. Cohen, I. M. Pike, D. G. Adler, M. B. Fennerty, I. Lieb, John G., W. G. Park, M. K. Rizk, M. S. Sawhney, N. J. Shaheen, S. Wani, D. S. Weinberg, Quality indicators for colonoscopy, Gastrointestinal Endoscopy 81 (2015) 31–53. URL: https://doi.org/10.1016/j.gie.2014.07.058. doi:10.1016/j.gie.2014.07.058.

[2] S. A. Hicks, A. Storås, P. Halvorsen, T. de Lange, M. A. Riegler, V. Thambawita, Overview of imageclef medical 2023 – medical visual question answering for gastrointestinal tract, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[3] B. Ionescu, H. Müller, A. Drăgulinescu, W. Yim, A. Ben Abacha, N. Snider, G. Adams, J. R. Meliha Yetisgen, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, N. Papachrysos, J. Schöler, D. Jha, A. Andrei, A. Radzhabov, I. Coman, V. Kovalev, A. Stan, G. Ioannidis, H. Manguinhas, L. Ştefan, M. G. Constantin, M. Dogariu, J. Deshayes, A. Popescu, Overview of imageclef 2023: Multimedia retrieval in medical, socialmedia and recommender systems applications, in: Experimental IR Meets Multilin-

guality, Multimodality, and Interaction, Proceedings of the 14th International Conference of the CLEF Association (CLEF 2023), Springer Lecture Notes in Computer Science LNCS, Thessaloniki, Greece, 2023.

[4] C. Hassan, M. Spadaccini, A. Iannone, R. Maselli, M. Jovani, V. T. Chandrasekar, G. Antonelli, H. Yu, M. Areia, M. Dinis-Ribeiro, et al., Performance of artificial intelligence in colonoscopy for adenoma and polyp detection: a systematic review and meta-analysis, Gastrointestinal endoscopy 93 (2021) 77–85.

[5] J. Y. Lee, J. Jeong, E. M. Song, C. Ha, H. J. Lee, J. E. Koo, D.-H. Yang, N. Kim, J.-S. Byeon, Real-time detection of colon polyps during colonoscopy using deep learning: systematic validation with four independent datasets, Scientific reports 10 (2020) 8379.

[6] A. Sánchez-González, B. G.-Z. Soto, Colonoscopy image pre-processing for the development of computer-aided diagnostic tools, in: S. Küçük (Ed.), Surgical Robotics, IntechOpen, Rijeka, 2017. URL: https://doi.org/10.5772/67842. doi:10.5772/67842.

[7] M. Soeder, A. Turshudzhyan, L. Rosenberg, M. Tadros, High-quality colonoscopy: A review of quality indicators and best practices, Gastroenterology Insights 13 (2022) 162–172.

[8] J. Bernal, J. Sánchez, F. Vilariño, Impact of image preprocessing methods on polyp localization in colonoscopy frames, in: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2013, pp. 7350–7354. doi:10.1109/EMBC.2013.6611256.

[9] M. Arnold, A. Ghosh, S. Ameling, G. Lacey, Automatic segmentation and inpainting of specular highlights for endoscopic imaging, EURASIP Journal on Image and Video Processing 2010 (2010) 1–12.

[10] S. A. Hasan, Y. Ling, O. Farri, J. Liu, H. Müller, M. P. Lungren, Overview of imageclef 2018 medical domain visual question answering task., in: CLEF (Working Notes), 2018.

[11] J. J. Lau, S. Gayen, A. Ben Abacha, D. Demner-Fushman, A dataset of clinically generated visual questions and answers about radiology images, Scientific data 5 (2018) 1–10.

[12] A. B. Abacha, S. A. Hasan, V. V. Datla, J. Liu, D. Demner-Fushman, H. Müller, Vqa-med: Overview of the medical visual question answering task at imageclef 2019., CLEF (working notes) 2 (2019).

[13] O. Kovaleva, C. Shivade, S. Kashyap, K. Kanjaria, J. Wu, D. Ballah, A. Coy, A. Karargyris, Y. Guo, D. B. Beymer, et al., Towards visual dialog for radiology, in: Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing, 2020, pp. 60–69.

[14] X. He, Y. Zhang, L. Mou, E. Xing, P. Xie, Pathvqa: 30000+ questions for medical visual question answering, arXiv preprint arXiv:2003.10286 (2020).

[15] A. Ben Abacha, M. Sarrouti, D. Demner-Fushman, S. A. Hasan, H. Müller, Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain, in: Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum-working notes, 21-24 September 2021, 2021.

[16] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, X.-M. Wu, Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering, in: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), IEEE, 2021, pp. 1650–1654.

[17] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[18] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).
[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, International journal of computer vision 115 (2015) 211–252.
[20] J. D. M.-W. C. Kenton, L. K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of naacL-HLT, volume 1, 2019, p. 2.
[21] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240.
[22] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.
[23] M. Schuster, K. K. Paliwal, Bidirectional recurrent neural networks, IEEE transactions on Signal Processing 45 (1997) 2673–2681.
[24] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using rnn encoder-decoder for statistical machine translation, arXiv preprint arXiv:1406.1078 (2014).
[25] Z. Yang, X. He, J. Gao, L. Deng, A. Smola, Stacked attention networks for image question answering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 21–29.
[26] J. Sacramento, R. Ponte Costa, Y. Bengio, W. Senn, Dendritic cortical microcircuits approximate the backpropagation algorithm, Advances in neural information processing systems 31 (2018).
[27] J. Lu, J. Yang, D. Batra, D. Parikh, Hierarchical question-image co-attention for visual question answering, Advances in neural information processing systems 29 (2016).
[28] M. Lin, Q. Chen, S. Yan, Network in network, arXiv preprint arXiv:1312.4400 (2013).
[29] H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T. D. Nguyen, D. Johansen, C. Griwodz, H. K. Stensland, E. Garcia-Ceja, P. T. Schmidt, H. L. Hammer, M. A. Riegler, P. Halvorsen, T. de Lange, HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy, Scientific Data 7 (2020) 283. URL: https://doi.org/10.1038/s41597-020-00622-y. doi:10.1038/s41597-020-00622-y.
[30] B. Iglewicz, D. Hoaglin, Volume 16: How to detect and handle outliers, The ASQC Basic References in Quality Control: Statistical Techniques, Edward F. Mykytka, Ph.D., Editor. 16 (1993).
[31] A. Telea, An image inpainting technique based on the fast marching method, Journal of Graphics Tools 9 (2004). doi:10.1080/10867651.2004.10487596.
[32] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.
[33] C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17, AAAI Press, 2017, p. 4278–4284.
[34] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L. Chen, Mobilenetv2: Inverted residuals

and linear bottlenecks, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA, 2018, pp. 4510–4520. URL: https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00474. doi:10.1109/CVPR.2018.00474.

[35] M. Tan, Q. V. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, 2020. arXiv:1905.11946.

[36] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, ICLR (2021).

[37] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jegou, Training data-efficient image transformers & distillation through attention, in: International Conference on Machine Learning, volume 139, 2021, pp. 10347–10357.

[38] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, 2021. arXiv:2103.14030.

[39] H. Bao, L. Dong, S. Piao, F. Wei, Beit: Bert pre-training of image transformers, 2022. arXiv:2106.08254.

[40] Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[41] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, CoRR abs/1505.04597 (2015). URL: http://arxiv.org/abs/1505.04597. arXiv:1505.04597.

[42] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. D. Lange, P. Halvorsen, H. D. Johansen, Resunet++: An advanced architecture for medical image segmentation, in: 2019 IEEE International Symposium on Multimedia (ISM), 2019, pp. 225–2255. doi:10.1109/ISM46123.2019.00049.

[43] D. Jha, S. Ali, K. Emanuelsen, S. A. Hicks, V. Thambawita, E. Garcia-Ceja, M. A. Riegler, T. de Lange, P. T. Schmidt, H. D. Johansen, D. Johansen, P. Halvorsen, Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy, in: MultiMedia Modeling, Springer International Publishing, Cham, 2021, pp. 218–229.

[44] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, H. D. Johansen, Kvasir-seg: A segmented polyp dataset, in: Y. M. Ro, W.-H. Cheng, J. Kim, W.-T. Chu, P. Cui, J.-W. Choi, M.-C. Hu, W. De Neve (Eds.), MultiMedia Modeling, Springer International Publishing, Cham, 2020, pp. 451–462.