Building An Automated Module for Image Quality Assessing from Narrow-Banding-Imaging Endoscopy Cameras

Pham Khac Long

Nguyen Thuan Thanh

Nguyen The Anh

Supervisor: Dr. Bui Van Hieu

Bachelor of Artificial Intelligence Hoa Lac campus - FPT University 2023

Acknowledgment

We would like to express our deepest gratitude to our supervisor, Dr. Bui Van Hieu for his precious guidance and helping us to complete this project.

Special thanks Dr. Trinh Dinh Hoan, Technological Expert at Viettel Cyberspace Center for providing dataset for our project.

We had the pleasure of working with Dr. Pham Quoc Dat, Department of Endoscopy and Functional Exploration, K Hospital, Tan Trieu Campus, who provided us with the understanding necessary background information.

Lastly, we also appreciate our families, FPT University and friends for always loving, support, encourage and give us the best through our work to finish this thesis.

Table of Contents:

List of Figures:

List of Tables:

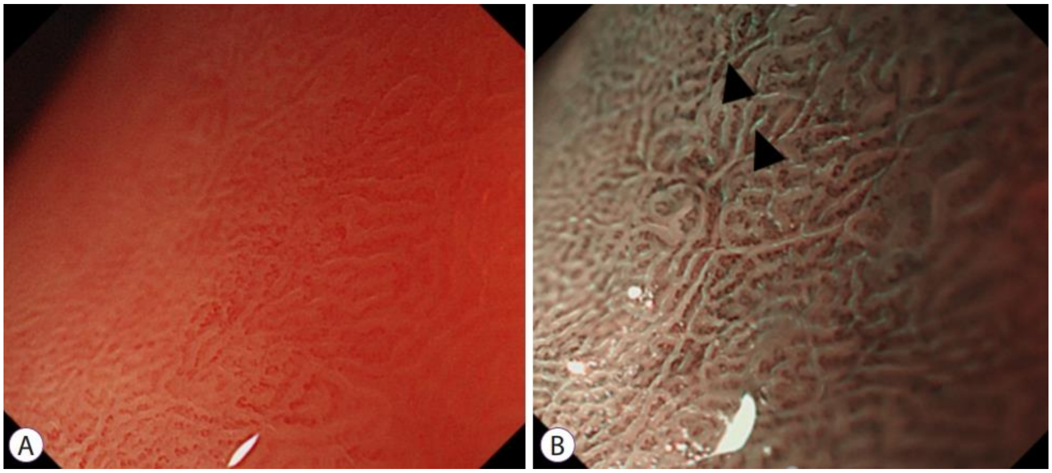
Abstract:

Gastrointestinal diseases have a significant impact on human well-being, and among them, gastrointestinal tumors have particularly high rates of occurrence and death. Besides traditional endoscopy, with the advancements in technology, endoscopic techniques nowadays such as the Narrow-banding-imaging (NBI) technique play a fundamental role in the quality of life for individuals by providing numerous imaging information for accurate diagnosis of gastrointestinal diseases. Although the amount of imaging data is abundant, the majority of them are substandard, making it extremely difficult for physicians to find quality images, therefore it is necessary to build an image quality control module for data cleaning. Based on the characteristics of the gastroscopy environment and inspired by non-reference image quality assessment (NR-IQA) methods, in this work, we utilized an IQA framework that assesses image quality from NBI endoscopy cameras that composes two stages utilizing deep learning approaches with some improvements. The first stage is based on a patch-based classification model to assess the quality of multiple regions of the image, which uses local features extracted from multi-layer of the convolutional neural network (CNN). In the second stage, these results of patches will be aggregated to give the final quality level for the entire image. Our improved pipeline shows over 96% and 97% with respect to overall precision and recall respectively. In addition, the final quality level of the endoscopic images was clinically evaluated by medical experts from Viet Duc and K Tan Trieu hospital. The results show that output quality levels are highly correlated to medical professionals' perception of endoscopic image quality. In terms of storage, the IQA module not only reduces nearly 90% the amount of storage capacity needed but also provides users a lot of flexibility depending on different scenarios. Finally, with the inference speed improvement method, we achieved 48 FPS in terms of frame rates, which is 4 times faster than the original.

Keywords: Image quality assessment, narrow-banding-imaging, endoscopy, deep learning, automated module.

1. Introduction:

Gastrointestinal diseases frequently occur in tropical countries and are often characterized by symptoms such as diarrhea, abdominal pain, abdominal distention, gastrointestinal bleeding, intestinal obstruction, malabsorption, or malnutrition. FGID, which stands for Functional Gastrointestinal Disorders, is commonly linked to chronic pain conditions like fibromyalgia as well as other functional syndromes like chronic fatigue syndrome. Moreover, two-thirds of individuals with FGID also experience psychological issues such as anxiety and depression [1]**.** Consequently, it is understandable that individuals with these conditions experience a significantly diminished quality of life, which may be even worse than those suffering from other chronic medical conditions such as severe congestive cardiac failure (grade III) or rheumatoid arthritis [2]**.** Gastrointestinal diseases are highly prevalent globally, affecting approximately 40% of the population. These conditions are more commonly observed in women compared to men and tend to decrease in frequency with age [3]. In primary healthcare facilities, gastrointestinal diseases contribute to around 12% of the overall patients, while constituting about 30% of the outpatient consultations in gastroenterology [4, 5]**.** Over 66% of individuals diagnosed with FGID have sought medical assistance within the past year, and approximately 40% of them rely on regular medication [6]. From an economic perspective, gastrointestinal diseases impose a substantial financial burden. In the year 2014 and 2015, treating these conditions cost the National Health Service (NHS) a minimum of £72.3 million. Out of this amount, two-thirds were allocated to expenses related to prescriptions, community care, and hospital treatments [7]. The inadequate quality of life experienced by individuals grappling with gastrointestinal disorders is a matter of great concern. If these conditions are not promptly recognized and addressed through a well-defined scientific treatment regimen, there is a substantial risk that they may escalate in severity. Left unchecked, such disorders could potentially advance to the point of developing into gastrointestinal cancers, further exacerbating the physical, emotional, and financial toll on patients while also increasing healthcare expenses. It is imperative that these issues be acknowledged and managed effectively to prevent their progression to more critical stages. Gastrointestinal cancers encompass a range of tumor types affecting various organs including the colon, rectum, stomach, pancreas, esophagus, anus, gallbladder, liver, and bile duct. Among gastrointestinal cancers, the most prevalent in the United States and many Western countries are colorectal cancer, stomach cancer, and pancreatic cancer. In fact, according to Global Cancer Statistics, during 2020, stomach cancer was identified as the fifth most diagnosed cancer worldwide and ranked as the fourth leading cause of death related to cancer [8]. In Vietnam, diseases of the gastrointestinal tract are the most common. During the Bridging Basic and Clinical Science for Gut Health conference, experts highlighted that 70 percent of the Vietnamese population is at risk of gastrointestinal diseases due to Helicobacter pylori (HP) infection [9]. These diseases, including constipation, digestive disorders, reflux, belching, and abdominal distension, already affect nearly 10 percent of the population. Moreover, more severe conditions such as cancer, peptic ulcer, irritable bowel syndrome, and enteritis are prevalent. Gastric cancer ranks as the country's third most common cancer, following liver and lung cancers. According to the Global Cancer Organization (GLOBOCAN), Vietnam witnessed over 14,000 new cases of colorectal cancer and more than 7000 deaths from it in 2018 [9]. The compromised quality of life endured by those afflicted with gastrointestinal disorders is indeed a pressing issue. Without swift recognition and intervention via a precisely outlined scientific treatment plan, there looms a significant peril of these conditions intensifying in seriousness. If allowed to fester, these disorders have the potential to evolve into full-fledged gastrointestinal cancers, compounding the already substantial physical, emotional, and financial burdens on patients. This, in turn, contributes to an escalation in healthcare costs. It is of paramount importance to duly recognize and adeptly address these concerns to avert their progression towards more critical phases**.** Gastrointestinal cancers have received extensive research attention and molecular characterization over the past two decades, particularly among solid cancers. Despite significant progress in understanding the molecular mechanisms behind these cancers, it is paradoxical that they continue to rank among the leading causes of death from cancer in Western countries. The initial stages of gastrointestinal cancers often present with non-specific symptoms, making early diagnosis challenging. As a result, many cases are diagnosed at advanced stages, where symptoms like bleeding, pain, or obstructions have already manifested, leading to 5-year survival rates below 30% [10, 11]. Therefore, the early detection of metastatic spread in gastrointestinal malignancies and early lesions in digestive tract diseases is crucial for stratifying patients for aggressive surgical interventions and guiding the use of additional therapies.



(a) (b)

Figure 1. Differences in images obtained from endoscopic images: (a) White light endoscopic image; (b) NBI endoscopic image.

To screen and track the spread of pathological gastroenterology signs, endoscopic imaging is the main method. White light endoscopy (WLE) is the initial method used for examining the gastrointestinal tract and remains an essential step in clinical examinations. However, WLE alone is not sufficient for accurately diagnosing gastrointestinal diseases due to its poor correlation with histopathological diagnosis. Moreover, WLE is prone to misdiagnosis, particularly in detecting early lesions of digestive diseases, as clear endoscopic visualization may be lacking. Therefore, diagnostic accuracy primarily relies on the expertise of the endoscopist [12]. In recent years, advanced endoscopic techniques have emerged to enhance detection accuracy. Narrow banding imaging (NBI) is one such method that improves the contrast between capillaries and submucosal vessels by manipulating the light source through specialized filters. By using narrow bands of light at specific wavelengths (415 nm - blue and 540 nm – green [13, 14]), NBI enables better visualization of vascular structures. The absorption peak of hemoglobin occurs at these wavelengths, which darkens the appearance of blood vessels and enhances their visibility. This facilitates the identification of other surface structures and enables observation of diseases associated with microvasculature. NBI has gained wide usage in the detection of gastrointestinal diseases. It allows for preliminary histological diagnosis of early esophageal lesions and assists in guiding targeted biopsies of lesions. Compared to WLE, NBI provides clearer images of capillaries or lesions and displays deeper structures under the mucosa of the digestive system, resulting in a higher detection rate for intestinal metaplasia of local lesions in the stomach [15].

However, despite many advancements of technologies and the abundance of medical images data from gastroscopy, endoscopic methods still have many challenges, for instance, image quality control is a key issue. Image quality can be negatively impacted or degraded for a variety of reasons. Some of them are related to clinical or anatomical aspects, such as the visibility of the transformation zone where gastric precancerous lesions tend to arise, and the presence of occlusion from vaginal tissue, blood, mucus and medical devices like a speculum, cotton, swab, or intrauterine device [16]. On the other hand, some of these including blur, noise, low contrast, etc. are related to the technical aspects of imaging instruments and lighting conditions, which depend heavily on the specific environment [17]. While it is crucial to train medical technologist how to capture high-quality photos, it is also pivotal to provide automated methods for limiting, controlling, and eliminating the image quality issue both in already-existing datasets and during acquisition. Specifically, in gastroscopy, each case usually takes approximately 15 to 45 minutes [18], with high resolution to serve the observation and diagnosis of the technician. In most cases, for the purpose of reviewing and subsequent treatment, each endoscopic video is recorded and stored as a series of image frames or a video in Picture Archiving and Communication system (PACS). Besides that, the recordings are supporting materials for explanations to the patients and valuable sources of information to be utilized for training young surgeons [19]. Due to the very high-resolution endoscopic videos, this procedure requires an extremely large and dedicated storage facility. In addition, the percentage of satisfactory frames accounts for a very small percentage of the entire video because the endoscopic video usually contains a lot of blurred and noisy frames. There are two fundamental reasons for this. First, the endoscopist moves the endoscope with his hands, which causes blurred and unstable frames because the lens of the endoscope has a high zoom factor. Additionally, the limited and fixed focus of the endoscopic camera makes it difficult to see the area that is currently out of focus. Second, a lot of frames have noise from fluid flushing, blood draining or tissue being cut by the endoscopist. Therefore, storing the entire video during endoscopy is wasteful and unnecessary. Furthermore, it makes extremely difficult and time-consuming for healthcare professionals to extract high-quality frames. Therefore, image quality assessment (IQA) has a crucial function in providing precise information to healthcare professionals, allowing them to detect, prevent, and monitor diseases at an early stage [20].

In this paper, the main contribution of our work is: we utilized the two-stages approach for assessing image quality from NBI endoscopy cameras. Firstly, the entire image is divided into non-overlap patches and classified into four categories that largely affect the endoscopic image quality according to medical professionals: **Brightness, Darkness, Motion-blur** and **High-quality**. Secondly, the aggregation process based on the statistical method is implemented to output the final quality level for the image with respect to five levels: **Bad, Poor, Fair, Good** and **Excellent**. We show that our method not only performs impressive in terms of statistical views but also has a high correlation with medical professional's perception of endoscopic image quality.

2. Related Works:

The goal of IQA is using computational models to simulate Human Visual System (HVS) to predict the quality of an image. Current IQA methods can be categorized into two groups based on different scenarios and conditions [21]. Full-reference (FR) and Non-reference (NR) IQA.

2.1. Full-reference image quality assessment (FR-IQA):

FR-IQA methods require two types of input: distorted and reference images to estimate their perceptual similarity. FR-IQA methods can be further divided into two categories: conventional evaluation metrics and learning-based models. The conventional metrics are based on a set of prior knowledge related to the HVS's characteristics. However, as visual perception is a complex process, it is challenging to imitate the HVS with a few hand-crafted elements. In contrast, learning-based FR-IQA models generate many general features from training data using various kinds of deep networks without the assistance of experts. In terms of metrics for optimization, the most commonly and widely used are PSNR and SSIM [22].

For learning-based methods, with the advent of deep learning and CNNs for specific, DeepQA [23] utilized CNN-based models to regress the sensitivity map to the subjective score that was produced from distorted and error maps. Ding et al. [24] developed a Deep Image Structure and Texture Similarity metric (DISTS) based on an injective mapping function, taking advantage of SSIM-like structure and texture similarity measurements. Siamese-Difference neural networks equipped with spatial and channel-wise attention were proposed by Ayyoubzadeh et al. [25] to predict the quality score. Most computer vision tasks, including IQA, have relied heavily on CNNs in recent years.

Unfortunately, the CNN-based models are unable to fully utilize the information across all regions of image since their internal limitation in capturing global features and their severe locality bias [26]. Following the success of vision transformer (VIT) [27], which employs Transformers [28] to capture the global dependencies of features, considerable advancements in a variety of computer vision tasks have been made. For the quality prediction task in IQA, IQT [29] used the reference and distortion picture characteristics derived by CNNs as the input of Transformer. To resolve the incompatibility of various input image sizes during training and testing, MUSIQ [30] employed Transformer to encode three scales of distortion image features.

However, due to the difficulty of collecting information about the references image in most practical scenarios, especially in the medical imaging field, NR-IQA, also known as blind image quality assessment (BIQA) are more useful in most practical applications [31].

2.2. Non-reference image quality assessment (NR-IQA):

The goal of NR-IQA is to provide a solution when a reference image is not available. In recent years, image quality assessment problems are attracting more and more attention from researchers because they objectively estimate the perceived quality of images without requiring a reference image for comparison. Besides, due to the lack of reference images, NR-IQA is also more challenging compared to FR-IQA. Ma et al. [32] proposed a multi-stage network that trains in two stages for distortion classification and quality prediction. In order to compensate for the lack of a true reference, Hallucinated-IQA [33] proposed an NR-IQA method based on generative adversarial models. To estimate the quality score, they paired the information from the hallucinated reference with the distorted image. Zhu et al. [34] proposed a model to capture the prior knowledge that is shared among various distortion types by leveraging meta-learning. Su et al. [35] proposed a model that predicts image quality by extracting content features from the deep model at multiple scales. In Hyper-IQA [36], the features were divided into low-level and high-level features, with the latter being transformed to redirect the former. Yunhong Li et al. [37] propose an unsupervised deep clustering method for non-reference IQA. Their 13-layer network upgrades fully connected layers to generate adaptive-sized high-level features. They also introduce a contracted regular term and a contracted autoencoder into the clustering loss function to form a quality model reflecting the data's clustering structure. Xiangfei Kong et al. [38] present two novel non-reference image quality metrics suitable for state-of-the-art image and video denoising algorithms, enabling auto-denoising. The metrics are designed to handle homogeneous regions and highly structured regions independently. However, the stability of these metrics is limited to scenarios with relatively low noise levels. Xialei Liu et al. [39] tackle the issue of a restricted IQA dataset size by employing a Siamese Network, a neural network architecture, to rank images according to their image quality. They utilize synthetically generated distortions with known relative image quality to train the network. This strategy aims to enhance IQA performance despite the limitations imposed by a small dataset.

Although widely used because does not depend on the reference image, it is very challenging for NR-IQA methods to find the suitable metric for assessing image quality with respect to each specific environment and condition. This is equivalent to finding the salient features of the image that are related to the quality. Therefore, it is necessary for searching an appropriate approach to extract feature depend on each specific domain. In the environment of gastroscopy, the quality in different image regions is very different, which usually heavily depends on external conditions such as uneven lighting, the motion speed of cameras, etc. While seeking to enhance endoscopy images, Tomoya Sato [40] shows that endoscopic images often have uneven lighting because of the way the endoscope is shaped and how its light shines. This uneven lighting can make it hard to see things clearly, especially in far-off areas that may look darker. Meanwhile, the research by Eric W. et al. [41] provides an insightful demonstration of endoscopy operations, elucidating how the underlying operational principles can lead to issues of excessive darkness or brightness within certain areas of the captured imagery. Also, to improve endoscopic images, Wei Tan et al. [42] found that clinical endoscopic images frequently suffer from problems such as uneven lighting, loss of fine details, and reduced contrast. These issues arise due to various factors, including the influence of light sources, hardware components, and different setup configurations used during clinical image capture. In the effort to improve endoscopic images to aid doctors in examining, diagnosing, and treating issues in the digestive system (esophagus, stomach, and intestines), Yang et al. [43] identified two primary factors contributing to the decline in image quality: motion blurring and low image resolution. Based on these studies, it has become evident that key factors requiring attention in endoscopic imagery include too bright, too dark, and motion blur. Thus, imputing the whole image into the image quality assessment model may not be optimal.

2.3. Patch-based image quality assessment:

To be able to take advantage of the local features of each region on the image, instead of inputting the whole image in the quality prediction model, there several approaches that divide the image randomly or consecutively into small patches and then aggregate the features of those patches to output the final image quality result. Gao et al. [44] first measured the local similarities of the feature maps from VGGNet layers between the reference and distorted images. The final quality score is then calculated by adding up all the local similarities. Similarly, by utilizing the learnt perceptual image patch similarity (LPIPS) metric, Zhang el al. [45] suggested using it for FR-IQA and demonstrated that deep features generated by using pre-trained DNNs outperform earlier traditional metrics by significant margin. A general end-to-end deep neural network called WaDIQaM [46] is proposed for the cooperative learning of local quality and local weights.

Image quality classification is a variant of the IQA problem depending on different contexts and conditions. [47] utilize a pre-trained CNN model, and the Support Vector Machine (SVM) is trained as an image quality classifier with inputs that are normalized features retrieved by the CNN model. Golchubian et al. [48] proposed a new dataset and an end-to-end pipeline using the CNN approach to classify images into six categories: bad lighting, Gaussian blur, motion blur, JPEG 2000, white-noise, and high-quality reference images. In terms of medical images, in order to identify between retinal fundus images of high and low quality, [49] proposed an algorithm that combines unsupervised features from saliency maps and supervised features coming from CNNs, which then are fed to the SVM for classification. [50] dealing with image quality control by classifying images into four quality categories: unusable, unsatisfactory, limited, and evaluable in order to enhance the effectiveness of automated visual assessment (AVE) for cervical precancer screening. An algorithm that combines both unsupervised features from a saliency map and supervised features from convolutional neural networks (CNN) was proposed by [51, 52]. These features are then utilized to train the SVM with the goal of automatically distinguishing between high-quality and poor-quality retinal fundus images. The integration of unsupervised and supervised features in this approach allows for more effective and automated image quality detection in the context of retinal fundus images. In 2020, Wang et al. [53] proposed a two-stage strategy for assessing the liver MRI image based on a classification problem. Each patch determined by radiologist to be diagnostic or non-diagnostic is utilized to train a CNN for segmentation purposes. Therefore, this method can focus on the regions of interest (ROIs). Then another CNN is used to classify the quality of extracted patches. Finally, the quantity of non-diagnostic patches over the liver patches of the images is used to evaluate the overall image quality. With respect to endoscopy imaging, Schoeffmann et al. [54] statistical method mainly based on the characteristics of endoscopic video and practical requirements to extract keyframes. However, due to the hand-crafted features extracted from characteristics of endoscopic image frames, they have to set the threshold manually, thus may be not generalized enough.

Overall, with respect to NR-IQA methods, especially in medical imaging, depending on the specific environment and conditions, the metrics for measuring performance need to be set up specifically. In terms of medical imaging, most of NR-IQA methods are designed for MRI images, therefore it is necessary to propose a new approach for assessing quality for endoscopic images that generalized enough. With respect to endoscopic environment, as mentioned above, different image regions have different qualities. Therefore, common NR-IQA methods may not be suitable for assessing endoscopic images. Inspired by [53], we utilized a two-stage NR-IQA framework for quality image from NBI endoscopy cameras that utilize deep neural network (DNN) models. In addition, to better separate between two classes with respect to patch-based classification, we apply the feature magnitude loss function which introduced in [55]. Our improved model shows impressive results both in terms of experimental and practical environment. Furthermore, we also proposed the inference speed improvement method by leveraging the parallel computation mechanism of Graphics Processing Unit, which reduced inference time significantly compared to the original.

3. Methods:

Our implementation idea is to evaluate the quality of each frame based on the classification problem. However, with each frame, the quality of each image area can be very different (Figure 3), making it difficult to assess the quality of the entire frame. Therefore, instead of putting the entire image through the model, we break the image into small patches, each patch has size of 128x128, and then perform classification on each of those patches before aggregating to make a final assessment for the frame image. Figure 2 shows the general idea of implementing our solution. Each frame will be divided into 20 consecutive image patches, each image patch is classified into 4 groups of characteristics as above, these results will be aggregated to give an assessment of the quality of the frame in which class of the 5 classes. Quality: **Bad, Poor, Fair, Good, and Excellent**.

Overall, the whole quality classification framework consists of two key components: **Patch-based classification** and **Quality assessment**.

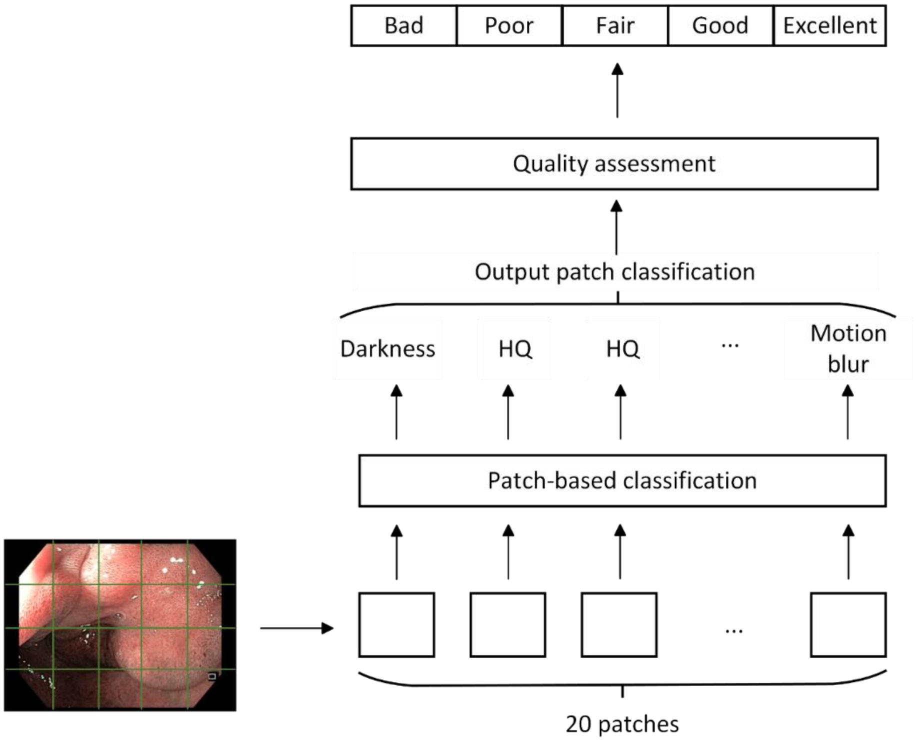


Figure 2. The diagram of the Endoscope Image Quality Assessment model.

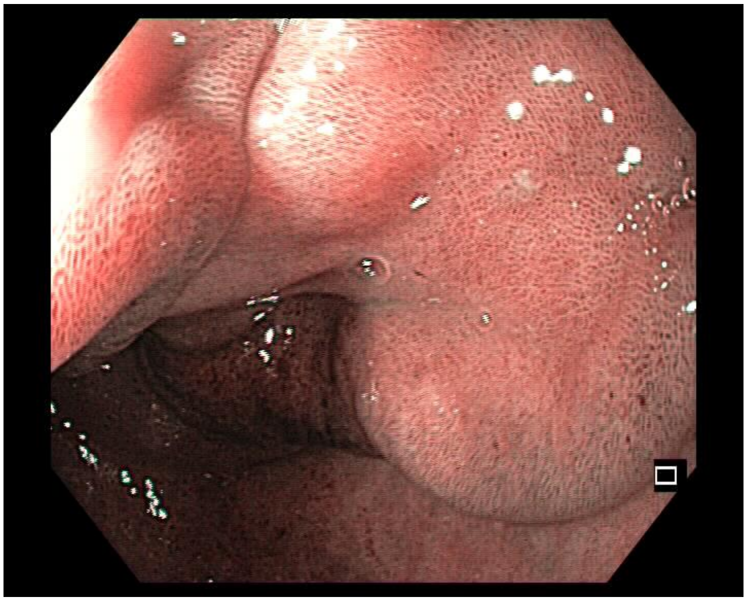


Figure 3. Describing uneven image quality across an endoscopic image.

While the upper left corner of the image is overexposed due to being too close to the light source, the lower left corner of the image appears very dark. The area of the image with good quality can be considered as the area that occupies about half of the image on the right because the texture on the surface can be clearly seen.

3.1. Convolutional Neural Network (CNNs):

Convolutional Neural Networks (CNNs) are a class of deep learning models primarily used for image recognition and computer vision tasks. They are inspired by the visual processing mechanism of the human brain and have proven to be highly effective in handling various image-related challenges. CNNs consist of multiple layers, including convolutional layers, pooling layers, and fully connected layers. The convolutional layers apply filters to the input image, extracting local features and detecting patterns such as edges, corners, and textures. The pooling layers reduce the spatial dimensions of the feature maps, preserving the most important information while reducing computational complexity. The fully connected layers process the high-level features learned by the previous layers, enabling the network to make predictions or classifications based on the input data. CNNs are trained using large labeled datasets, and the learning process involves adjusting the weights of the network to minimize the difference between predicted outputs and actual labels. This process is typically done through optimization techniques like gradient descent [56]. CNNs have achieved remarkable success in various computer vision tasks, such as image classification, object detection, semantic segmentation, and image generation. Due to their ability to automatically learn hierarchical representations from data, they have become the backbone of many state-of-the-art image processing and computer vision applications.

There are many pre-trained CNNs that come with various architectures, each having different internal layers and techniques. One notable example is GoogLeNet, which introduces the concept of Inception Modules. These modules use convolutional filters of different sizes and then concatenate the outputs from these different-sized filters to form the input for the next layer. The use of Inception Modules allows GoogLeNet to capture information at multiple scales and learn different features, making it an important and effective CNN armature for various computer vision tasks [57]. In contrast to GoogLeNet, AlexNet follows a different approach and does not use filter concatenation. Instead, it utilizes the output of the previous layer as the input for subsequent layers.

Mostly in order to solve a complex problem, increasing depth is the trending solution by stacking to layers in Deep Neural Networks. The intuition behind adding more layers is that these layers progressively learn more complex features. However, simply increasing the depth of the network causes the vanishing gradient, and Resnet [58] architecture emerged to solve this problem thoroughly.

3.2. Resnet18:

Nowadays, deep architectures are becoming trending cause a good approximation of the data is not only the goal, but we also need a model capable of generalizing the data. However, increasing the network depth is not just simply stacking layers together. When the size of the network increases, the vanishing gradient will happen and lead the network's performance to degrade rapidly [58]. To solve the degradation problem of training very deep networks, Kaiming He et al. developed ResNet or Residual Network, a deep learning network that has become very popular in the field of computer vision. ResNet makes it possible and efficient to train hundreds, even thousands of layers of neural networks while still ensuring good performance. ResNet solves this problem by adding Shortcuts, also known as Skip Connections, to traverse one or more layers [58]. Figure 4 illustrated the idea of skip connections.

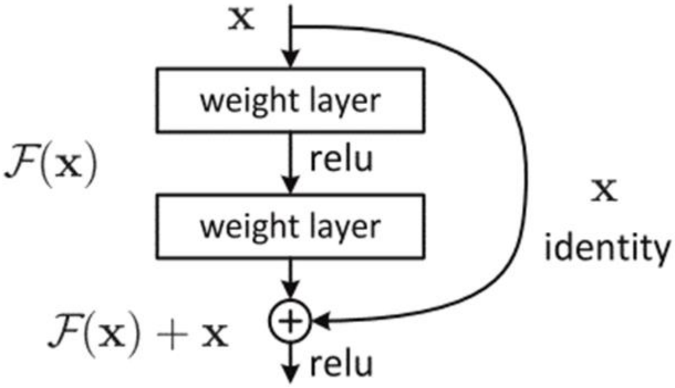
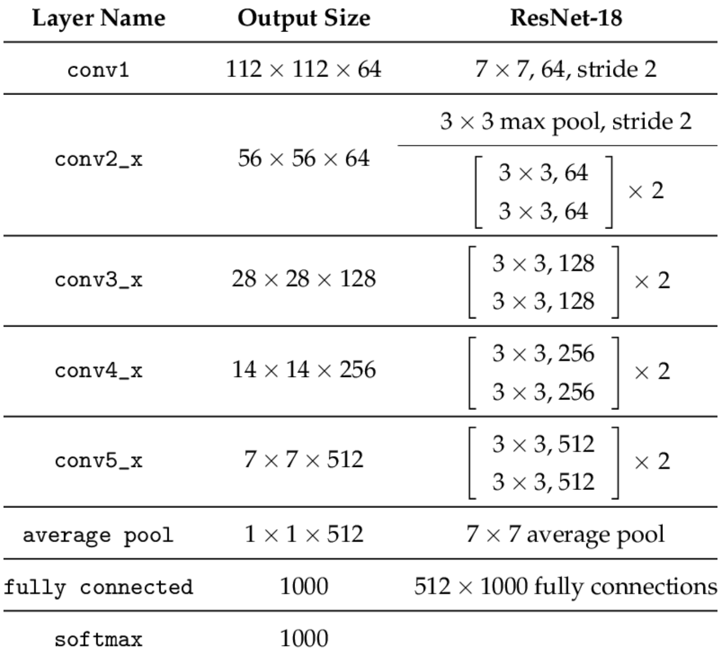

x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 4. The skip connections in ResNet [58]

The Skip Connections in between layers combine the outputs of earlier layers with the outputs of the layers that are stacked on top of them, and this makes the network's performance will not degrade while stacking layers. Moreover, with this architecture, the upper layers receive information in a more direct manner from the lower layers, resulting in a more efficient adjustment of weights. The manifold advantages demonstrated by ResNet in the realm of computer visual tasks have rendered it an invaluable asset for our research endeavors. Building upon this foundational framework, we shall endeavor to apply this neural network, fortified by a series of deliberate and refined structural enhancements, as a robust and versatile solution to effectively tackle the complex and multifaceted challenges that confront us in our research domain. In our training phase, we opted for the ResNet18 architecture to optimize processing time, a crucial factor in our network. The specific architecture details can be found in Figure 5 below.



Layer Name	Output Size	ResNet-18
conv1	$112 \times 112 \times 64$	7×7, 64, stride 2
conv2_x	$56 \times 56 \times 64$	3×3 max pool, stride 2 $\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$
conv3_x	$28 \times 28 \times 128$	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$
conv4_x	$14 \times 14 \times 256$	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$
conv5_x	$7 \times 7 \times 512$	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$
average pool	$1 \times 1 \times 512$	7×7 average pool
fully connected	1000	512×1000 fully connections
softmax	1000	

Figure 5. Resnet18 architecture [58]

With the widely use, showing impressive results in many competitions and publications, to balance performance and speed, we decided to choose ResNet18 as the patch classification model. In addition, we also make some minor architectural improvements to the model to make it work better for the IQA problem.

3.3. Implementation:

3.3.1. Patch-based classification model:

There are many factors that can adversely affect or degrade image quality, often depending heavily on the surrounding environment. In the domain of gastroscopy, according to [40,41,42,43] and the statistics of medical professionals, the quality of each frame is largely influenced by the following

4 factors: **Brightness**, **Darkness**, **Motion blur** and **High quality**. Therefore, patches are divided are classified into 4 main groups:

1. Brightness (Figure 6(a)) due to the light reflection of gastric juice and the light intensity of the camera.

2. Darkness (Figure 6(b)) because the lighting cannot be evenly distributed over the whole scene.

3. Motion blur (Figure 6(c)) due to the relative motion of the camera and the stomach wall surface. This factor accounts for a very large proportion of the endoscopic image.

4. High quality (Figure 6(d)): Sharp areas, usually only present when relative camera motion and stomach surface are low.

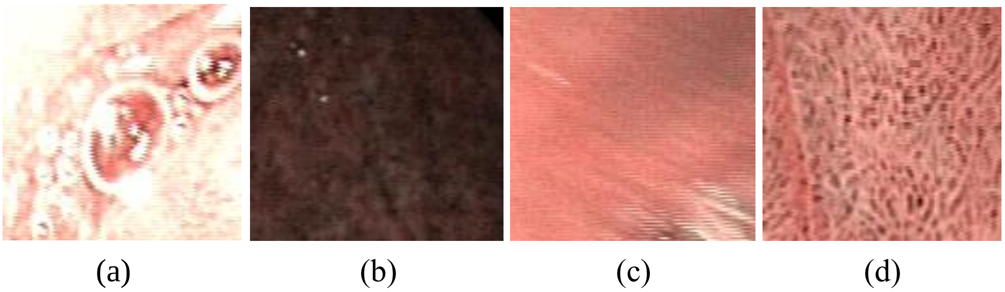


Figure 6. The key feature classes affecting NBI endoscopic image quality: (a) Brightness; (b) Darkness; (c) Motion blur; (d) High quality.

To classify the image quality for each small patch, we utilize deep learning models. Convolutional neural network (CNN) exploits the structure of the image through local interactions captured by convolutions with small filters. Therefore, different layers of the network may contain different semantic information, and the deeper into the network, the more complex features are obtained. Classification of image quality depends on both low-level features and high-level features of the image [27], in other words, the representation of a particular type of distortion may be different at each layer in the image in a deep learning network. Therefore, using only features in the final convolutional layer may not be enough to predict the quality of an image. Inspired by the hierarchical feature extraction strategy in [32], we propose a solution that combines both low-level as well as high-level representation at different layers of the network by resizing and concatenating them, then passing through Global Average Pooling (GAP) and a Fully connected layer (FC) for the purpose of returning to the desired size. The architecture diagram is depicted in Figure 7.

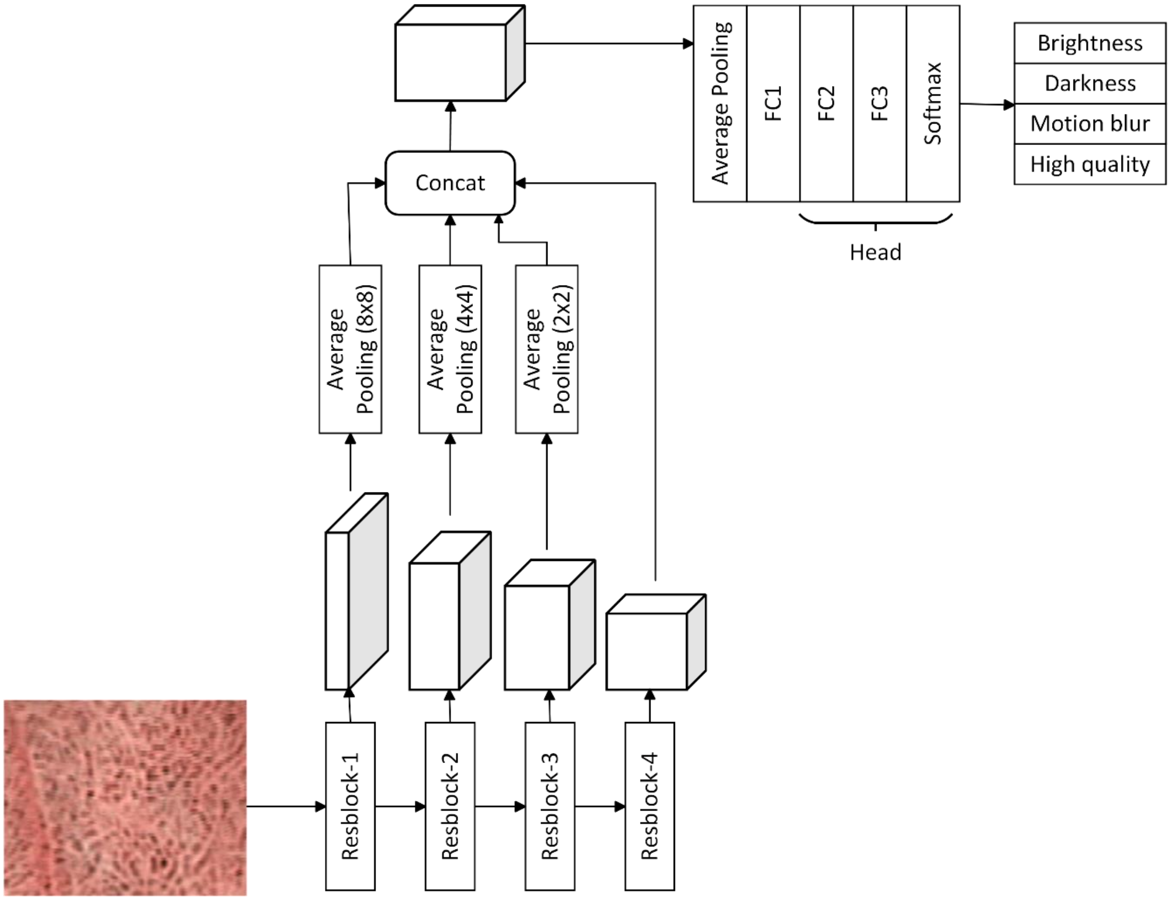
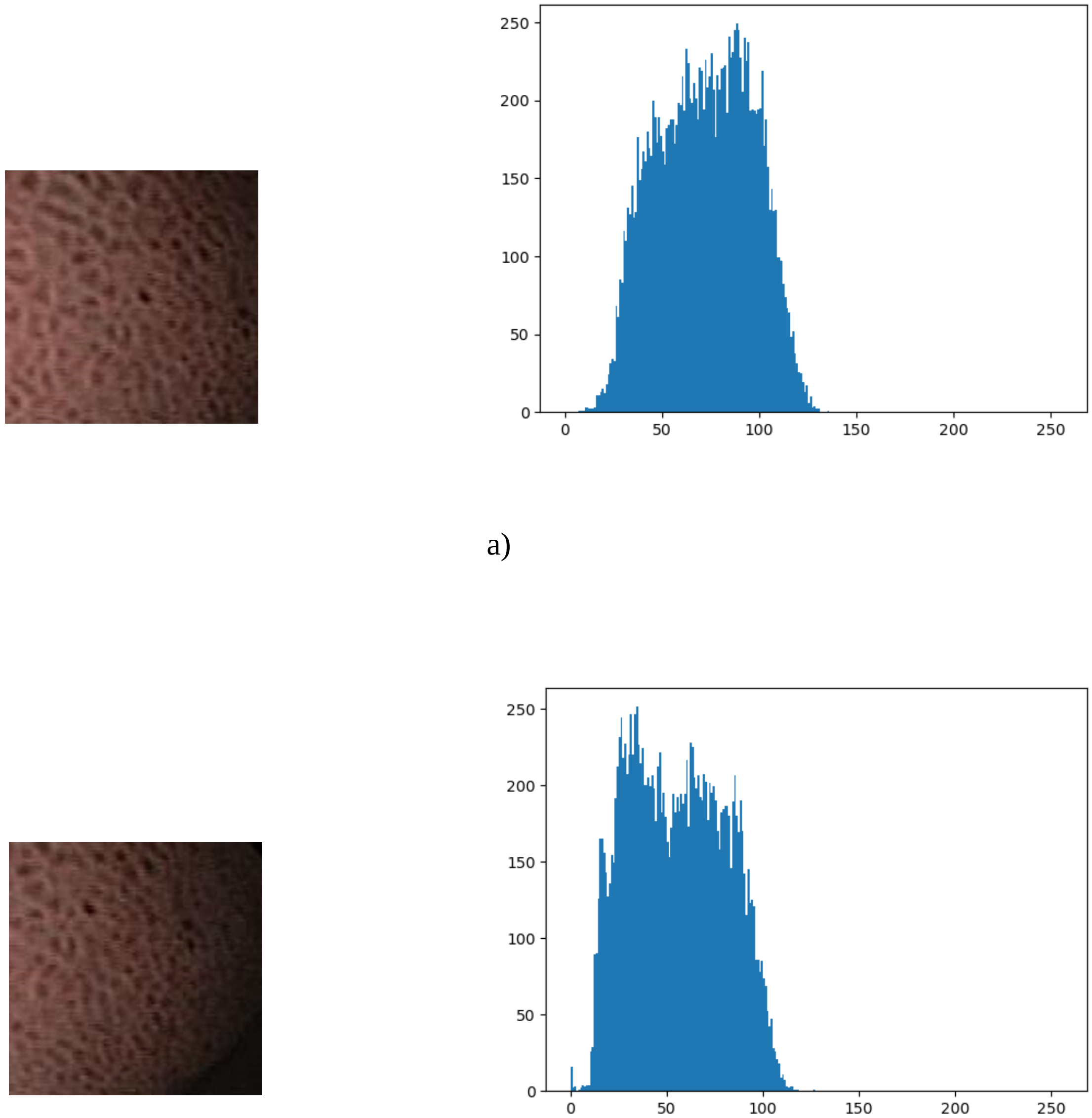


Figure 7. The improved AI model architecture for endoscope image patch quality classification.

As can be seen in Figure 8, the two patches of darkness and high-quality patch images are quite similar visually. As a result, this makes the model easy to confuse in distinguishing these two types of image patches. Therefore, to improve this problem, inspired by [55], which tried to separate the normal and abnormal video snippet by proposing the feature magnitude loss function, in addition to only using the loss function for the classification problem, we also use the loss function for the deep features of the above two patch types taken from the FC layer utilizing feature magnitudes. More specifically, we utilized the feature magnitudes loss that maximizes the separability between darkness and high-quality patches.



a)

b)

Figure 8. The relative similarity between high-quality and darkness patches: (a) High-quality patch and its corresponding histogram; (b) Darkness patch and its corresponding histogram.

Given the entire image $I = \{(P_i,\ y_i)\}_{i=1}^{N}$, where N the total number of patches in an image, P_i is the ith patch and y_i is the label. The improved network is denoted by $r_{\theta,\emptyset} = f_{\emptyset}(s_{\theta}(P_i))$ and return a N-dimensional feature $[0,1]^T$ representing the classification of all patches into 4 labels above, with the parameters $\theta, \emptyset$ defined below. The end-to-end model is training with the total loss:

$$l = min_{\theta,\emptyset} \sum_{i,j=1}^{N} (1-\alpha) l_s\big(s_\theta(P_i), s_\theta(P_j), y_i, y_j\big) + \alpha l_f\big(f_\emptyset\big(s_\theta(P_i)\big), y_i\big),$$

where $s_\theta: P \to X$ is the patch feature extractor (with $X \subset R^D$ is pre-computed feature of dimension D from an image patch), $f_\emptyset: X \to [0,1]^T$ is the patch classifier, α is assigned weight for each term, $l_s(.)$ denotes a loss function that maximises the separability between the darkness and high-quality patches, and l_f is a loss function to train the patch classifier. The feature magnitude loss function can be further defined as:

$$l_s\big(s_\theta(P_i), s_\theta(P_j), y_i, y_j\big) = \max\Big(0, m - d\Big(g_\theta(X_i), g_\theta(X_j)\Big)\Big)$$

$$if\ y_i, y_j \in \{Darkness, High - quality\}$$

where m is pre-defined margin, $X_i = s_\theta(P_i)$ is the patch feature, g_θ calculates the feature magnitude of the patch feature, and d represents separability function that computes the difference between two feature magnitudes.

3.3.2. Quality assessment:

Inspired by [44,45,46] in evaluating patch quality locally, then synthesizing and giving results globally, for each frame after scoring the quality for all patches, the aggregation process to evaluate the frame quality in 5 levels from low to high: Bad, Poor, Fair, Good and Excellent is performed statistically. After looking at the characteristics of Good and Excellent quality images as classified by medical professionals, we found that for these images, the high-quality patches are often located adjacent horizontally or vertically to each other, forming an area with a lot of medical information in it. Therefore, we decided to use the Breadth-first search (BFS) algorithm to search and based on the percentage of adjacent high-quality patches to evaluate the quality for the entire frame. The percentage of adjacent high-quality patches is defined as follows:

$$n = \frac{N}{total\ number\ of\ patches\ per\ image},$$

where n is the percentage of adjacent high-quality patches and N is total number of adjacent high-quality patches. Similarly, the percentage of patch type remaining is equal to the number of patches of that type divided by the total number of patches per image. In addition, we classified the images as Bad and Poor quality based on the percentage of motion blur patches because we found that in an endoscopic video, most of the very poor-quality frames were those that were blurred due to the relatively fast camera transitions, causing the loss of almost all areas containing medical information. Finally, images that are concluded to be Fair quality are classified in cases where the camera is stationary or has relatively slow motion, so blurred patches account for a small percentage. However, the light source emitted is not evenly distributed over the entire image, causing patches of brightness and darkness to dominate and only a small area of high-quality patches. Assuming the percentage of each patch type in an image is illustrated in Table 1 below:

Table 1. The percentage of each patch type in an image.

Brightness	Darkness	Motion blur	High-quality (adjacent patches)
a%	b%	c%	n%

Then, the quality rating for the entire frame is performed as shown in Table 2:

Table 2. Quality assessment for the entire frame.

Bad	Poor	Fair	Good	Excellent
c > 45%	35% ≤ c ≤ 45%	c < 35% and n < 35%	35% ≤ n ≤ 55% and c < 35%	n > 55% and c < 35%

In general, the factors to evaluate the quality of gastroscopy images are applied as follows:

- **Bad quality image (Bad)**: If the patch rate is blurred over 45%.

- **Poor quality image, difficult to find medical information (Poor)**: If the blur rate is from 35% to 45%.

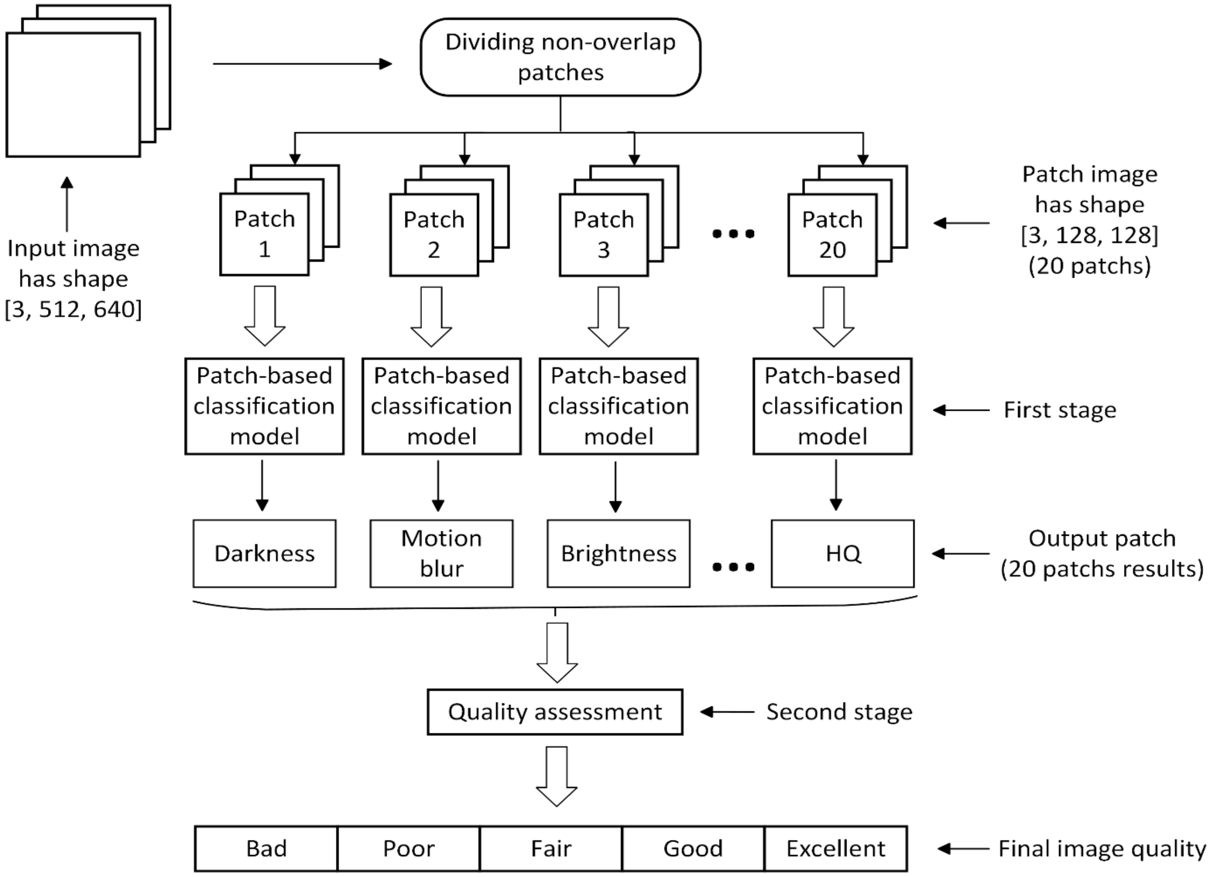
- **Acceptable image quality (Fair)**: If the patch ratio is blurred below 35% and the high-quality patch ratio is below 35%. This is the case when the camera has little movement (less blur) and a small number of image areas are still good enough to observe medical information well.

- **Good quality image (Good)**: If the patches rate of high-quality is from 35% to 55%.

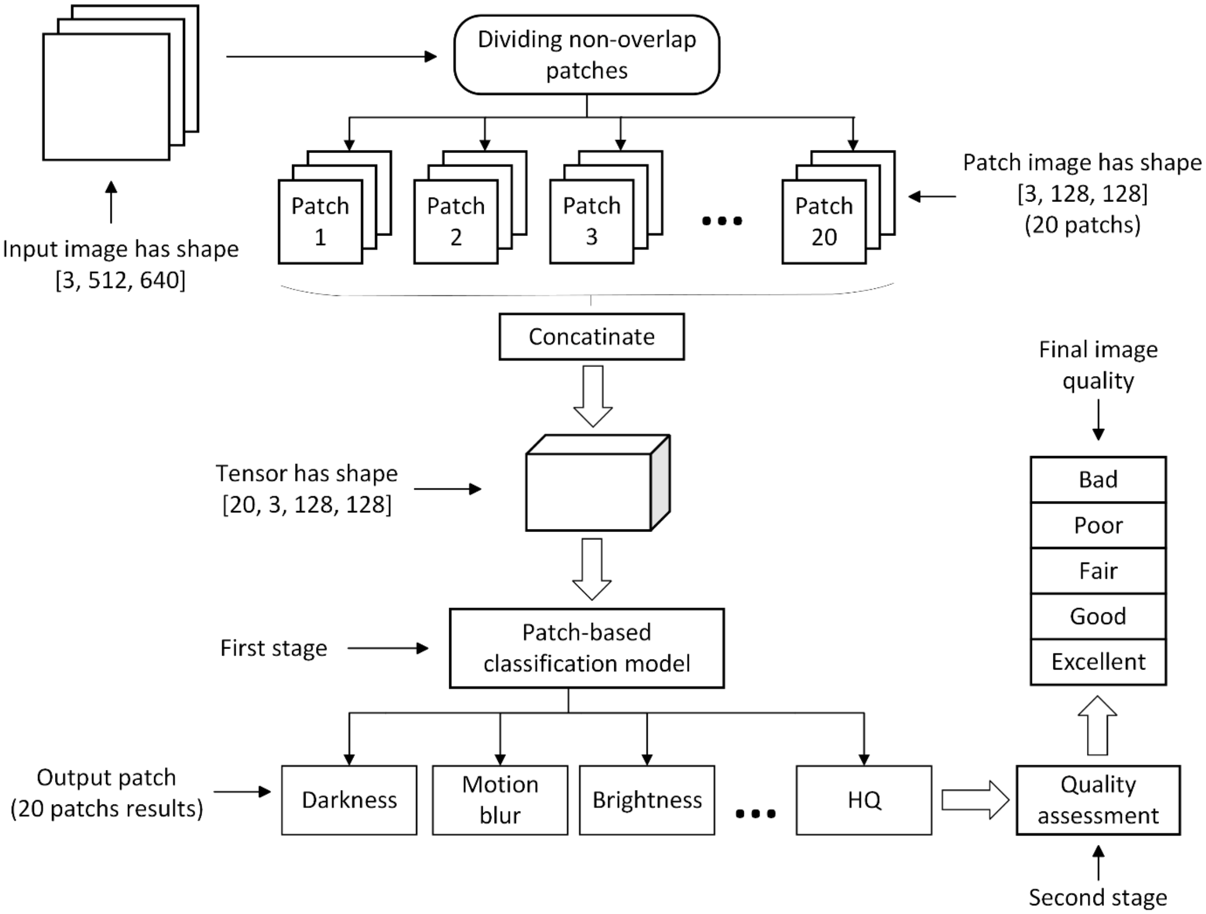
- **Very good quality image (Excellent)**: If the patches rate of high-quality accounts for 55% or higher.

3.3.3. Inference process:

To speed up the inference process, instead of classifying each image patch individually (Figure 9a), we leverage the power of the Graphics Processing Unit (GPU) by after dividing all patches, we concatenate all those patches into a single tensor in Pytorch [59] has size: [20, 3, 128, 128], where 20 is the total number of patches, 3 is the number of channels RGB and 128x128 is the patch size (Figure 9b).


Dividing non-overlap patches
Input image has shape [3, 512, 640]
Patch 1
Patch 2
Patch 3
Patch 20
Patch image has shape [3, 128, 128] (20 patchs)
Patch-based classification model
First stage
Darkness
Motion blur
Brightness
HQ
Output patch (20 patchs results)
Quality assessment
Second stage
Bad
Poor
Fair
Good
Excellent
Final image quality

(a)


Dividing non-overlap patches
Patch 1
Patch 2
Patch 3
Patch 20
Patch image has shape [3, 128, 128] (20 patchs)
Input image has shape [3, 512, 640]
Concatinate
Final image quality
Tensor has shape [20, 3, 128, 128]
Bad
Poor
Fair
Good
Excellent
First stage
Patch-based classification model
Output patch (20 patchs results)
Darkness
Motion blur
Brightness
HQ
Quality assessment
Second stage

(b)

Figure 9. The inference process: (a) Original way; (b) Proposed way.

With this improved inference method, the speed of the whole pipeline can be enhanced significantly. The time results measurement of two ways of implementing the inference process will be shown in detail in section 5.

4. Data Preparation:

The original dataset is the private dataset from the medical image database of Viettel Cyberspace Center. Specifically, the raw data is a set of endoscopic images extracted from specialized NBI cameras in real-world endoscopy cases, with an original size of 720 x 576 (width x height). Since the original image data has a black border surrounded (Figure 3), which is not necessary information, and to bring it to the valid size to be divisible by the 128x128 size of the patch, we cut off most of the surrounding black border, thus the size of images for training or testing process became 640x512.

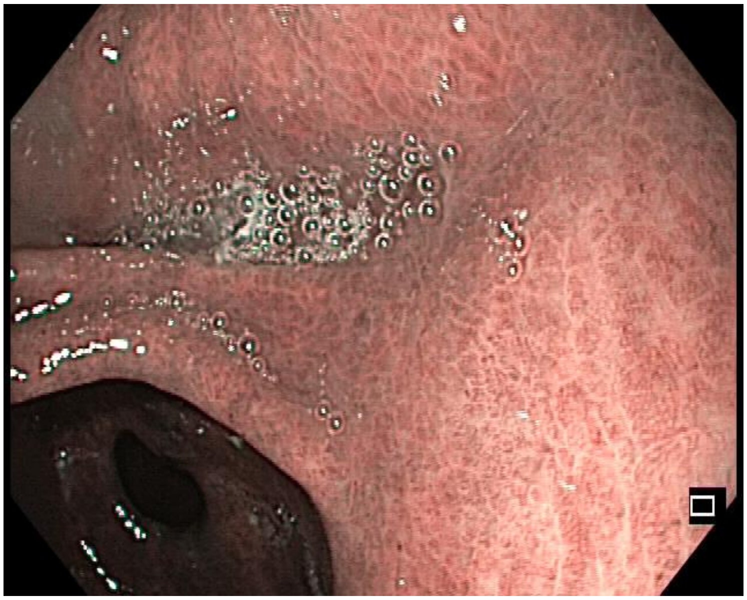


Figure 10. Endoscopic image after cutting off most of the black border has size of 640x512.

To be able to train the patch-based classification model, we first crop the image, then divide the image into consecutive patches, each patch is 128x128, and then divide it into 4 categories as described above under the guidance of medical professionals. Therefore, the entire image will have a total of 20 patches. The number of each type is described in Table 3 below:

Table 3. Dataset used for training patches-based classification model.

	Brightness	Darkness	High-quality	Motion blur
Train	205	592	3278	918
Val	107	212	699	639

Because the initial data is quite imbalanced when the number of patches with high-quality accounts for the majority, 6 times more than the number of patches that are darkness and 12 times more than the number of patches that are brightness. Therefore, before putting into the training model, we use data augmentation methods to bring the amount of data of each label to the balance level. Specifically, the data augmentation techniques that we take advantage of are mainly based on geometric transformations such as rotation, flip, crop, stretch, etc. and intensity transformations such as gamma correction, contrast change, etc. Finally, we normalize the input to the form (0,1) and then return it to mean = [0.485, 0.456, 0.406] and std = [0.229, 0.224, 0.225]. After applying data augmentation techniques, clearly, the number of each type became much more balanced as shown in Table 4:

Table 4. Dataset after using the data augmentation techniques.

	Brightness (x12)	Darkness (x5)	High-quality	Motion blur (x3)
Train	2460	2960	3278	2754
Val	107	212	699	639

Figure 11 depicts the number of training dataset before and after applying data augmentation techniques. Clearly, the augmented training dataset is much more balanced compared to the original.

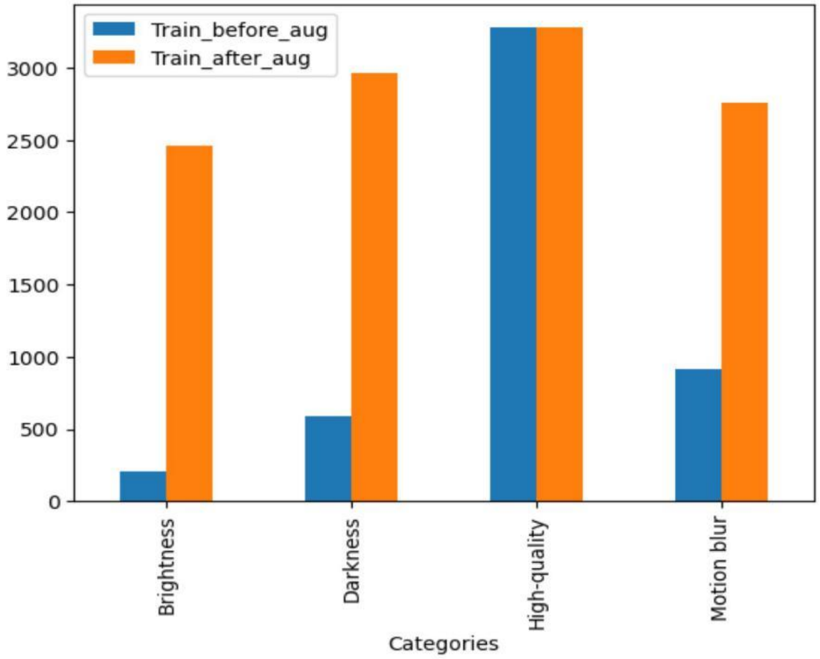


Figure 11. Demonstrate the quantity of training dataset before and after utilizing augmentation techniques.

5. Results:

5.1. Patch-based classification model:

During training, we chose Adam [60] as the optimizer with learning rate = 1e-4 and weight decay = 5e-3, batch size = 32. In addition, we also use the learning rate reduction strategy Cosine annealing restart to help improve the model's ability to pass local optimal points. Regarding the loss function for patch classifier, we choose Cross entropy loss with the hyperparameter $\alpha = 0.2$ the margin $m = 15$ with respect to feature magnitude learning function. Similar to other classification problems, we use measurement methods: accuracy, precision, recall and F1-score. The results of the baseline model classification (without using data augmentation techniques, multi-feature fusion strategy and feature magnitude learning) for patches are shown in Table 5 below:

Table 5. Experimental results of baseline patch-based classification model.

Class	Precision	Recall	F1-score	Accuracy
Brightness	88.23	98.13	92.92	95.65
Darkness	87.61	93.40	90.41	
Motion blur	97.53	98.75	98.13	
High-quality	97.89	93.13	95.45	

Patches are classified into 4 classes and evaluated according to 4 metrics: Precision, Recall, F1-score, and Accuracy. From Table 5, we can see that because of the numerical imbalance between classes, the results of the baseline approach are clearly unstable. Specifically, although the value of recall in the two groups brightness and darkness are significantly high, which are 98.13 and 93.40 respectively, the precision metric of those is extremely low that are below 89, resulting in a remarkably low F1-score. In addition, due to the relative similarity of the two types of patches that are darkness and high-quality as mentioned above, the precision of the darkness and the recall of the high-quality types are very low, which are 87.61 and 93.13 respectively. In contrast, by using data augmentation techniques, multi-feature fusion strategy and feature magnitude learning, the patch-based classification results of the improved version can be improved significantly as shown in Table 6:

Table 6. Experimental results of patch-based classification improved model.

Class	Precision	Recall	F1-score	Accuracy
Brightness	97.25	99.07	98.15	97.7
Darkness	**92.96**	**93.40**	**93.18**	
Motion blur	99.06	99.06	99.06	
High-quality	**97.84**	**97.42**	**97.63**	

By applying different data augmentation methods, the problem of data imbalance can be partially solved, thus significantly improving the results of the precision index in the two groups of brightness and darkness patches, which are 97.25 and 92.96 respectively. As a result, the F1-score value can be also enhanced to 98.15 and 93.18, showing the importance of numerically balancing between types of patches. Moreover, the confusion between the high-quality and darkness types can be improved by utilizing the feature magnitude loss function, which increases the recall of high-quality type to 97.42. In addition, the improved version has 97.7% with respect to overall accuracy, higher than 2% compared to the baseline. In general, with the application of the techniques mentioned above, the results of the patch-based classification model have improved significantly on almost all indicators.

Because of the extreme imbalance of the validation set, the overall accuracy may not be reflected enough in the effectiveness of the model. Therefore, we plot the F1-score of two version baseline and improved patch-based classification model according to four categories to demonstrate the enhancement. Figure 12 shows the F1-score of the baseline and improved model.

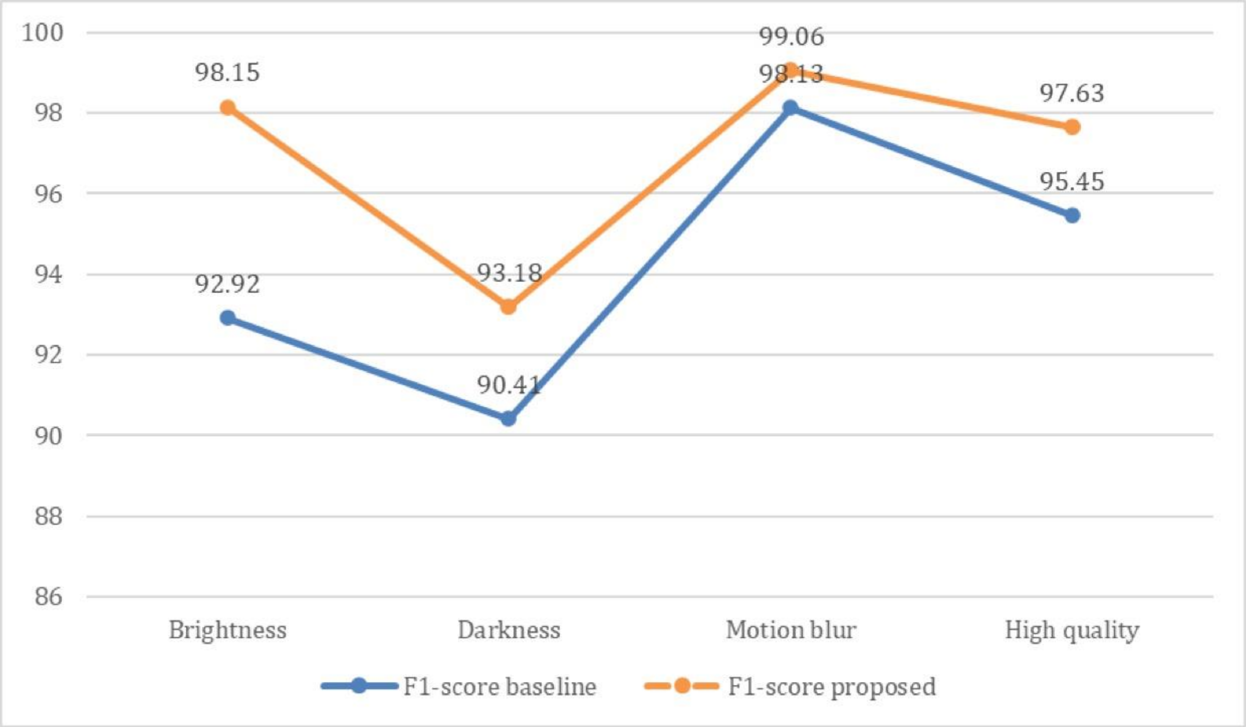


Figure 12. F1-score of the baseline and improved patch-based classification model.

Specifically, the F1-score of brightness and darkness types increase significantly from 92.92 to 98.15 and 90.41 to 93.18 respectively. With the rise of F1-score in terms of the brightness type, that of the high-quality type of also growths remarkably from 95.45 up to 97.63. From there, we can see significant improvements of the patch-based classification model through changes in model architecture as well as the loss function.

Figure 13 illustrates the t-SNE plot of the features extracted of improved Resnet18 model from validation set trained using baseline and improved approach. It shows clearly that the improved pipeline has better separation between classes than the original one, especially between high-quality and darkness types. The classification confusion matrix of two approaches calculated from the validation set are also in Figure 14 shown below.

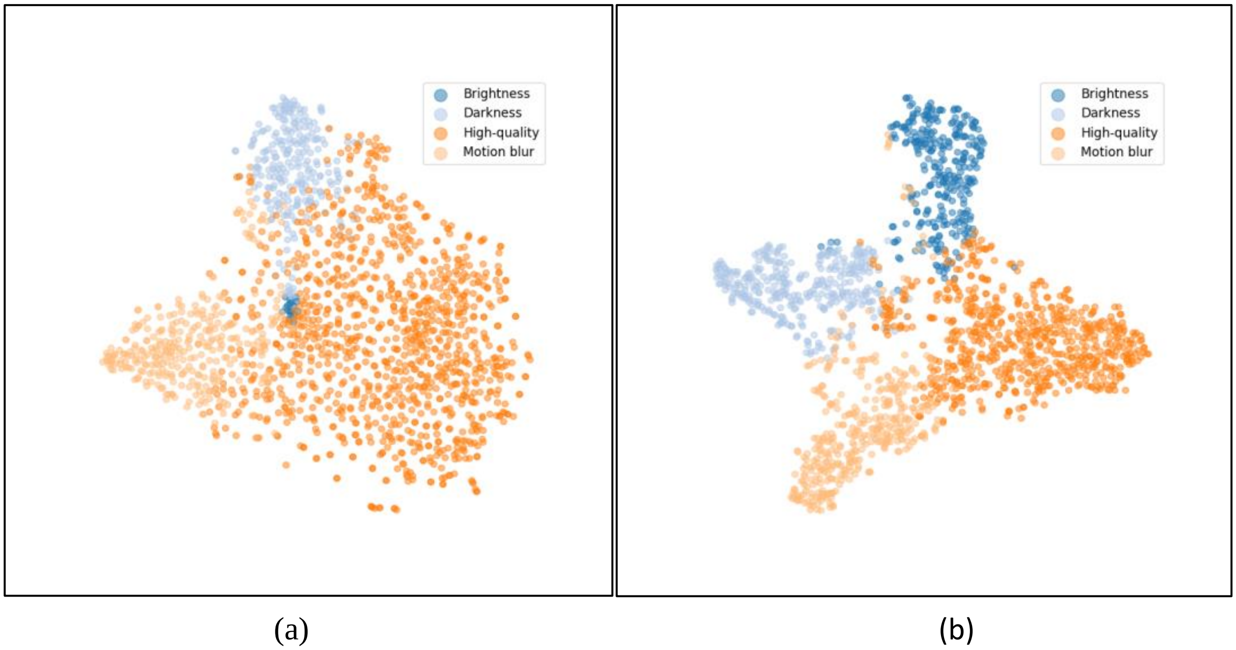
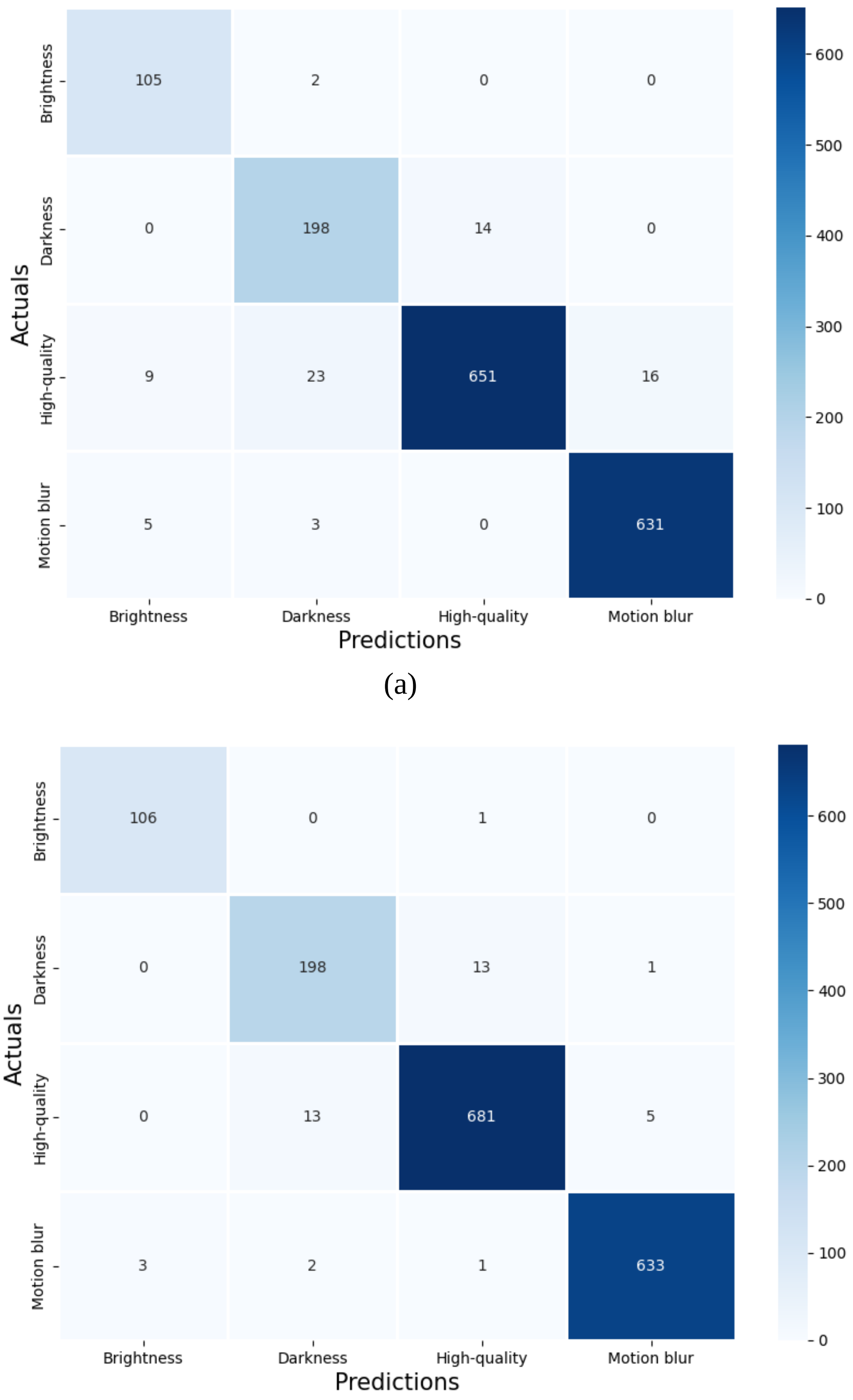


Figure 13. t-SNE plots of patch-based classification: (a) Baseline model; (b) Improved model.

From Figure 14, the high-quality type of prediction is most significantly improved. Specifically, the error prediction between brightness and high-quality types has been completely resolved by reducing the number of falsely predicted brightness patches from 9 samples to 0 samples. Especially, the confusion between brightness and high-quality types has also been significantly decreased, from 23 samples to 13 samples. In addition, the number of false motion blur predicted has decreased significantly from 16 samples down to 5 samples. Therefore, the number of correctly predicted high-quality patches was boosted remarkably from 651 to 681.



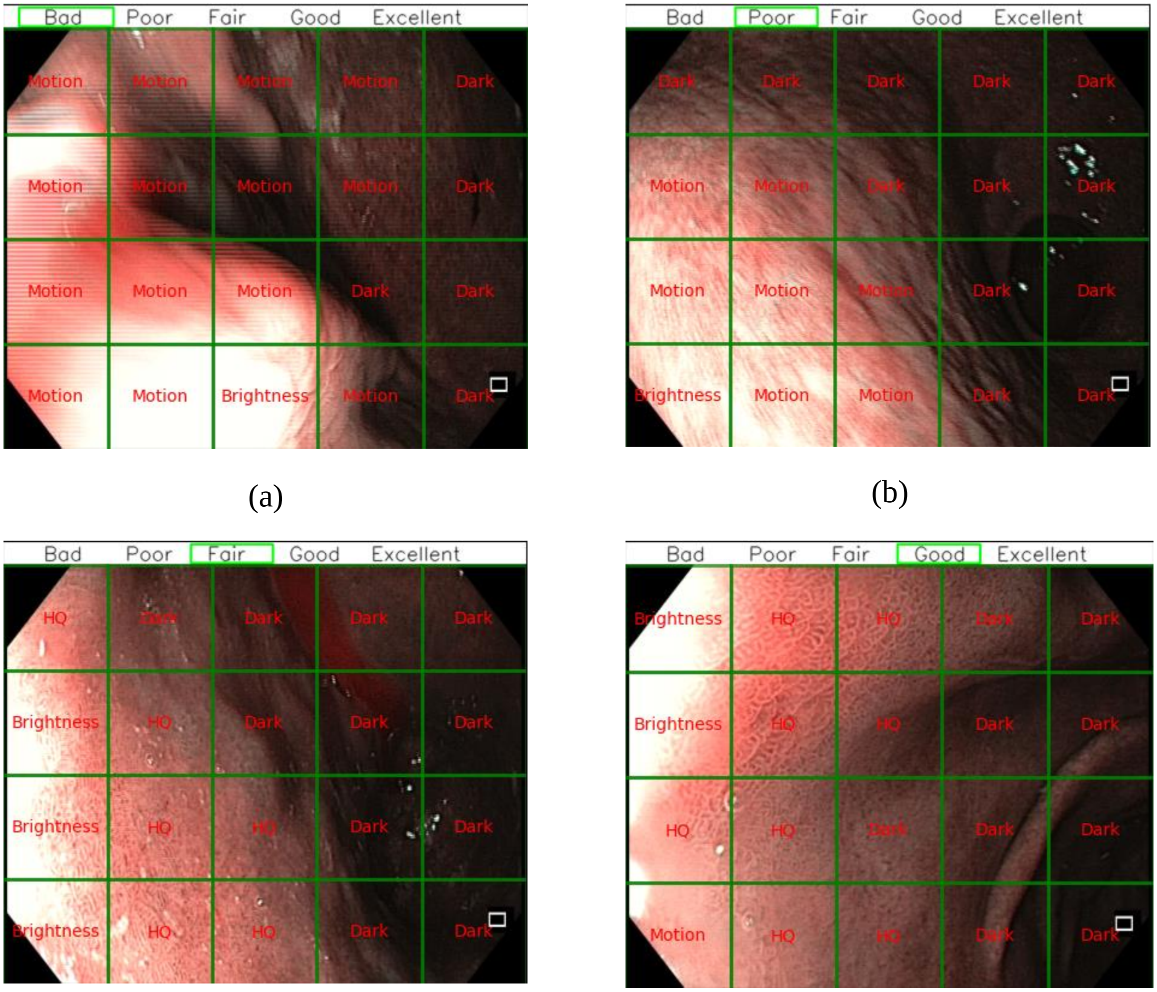
(a)

(b)

Figure 14. Confusion matrixes of patch-based classification model: (a) Baseline model; (b) Improved model.

5.2. Results in practical environment:

After building the model for patch classification, we calculate the number of each type and give an assessment for the entire frame. Figure 15 below shown a several test results on the practical environment:

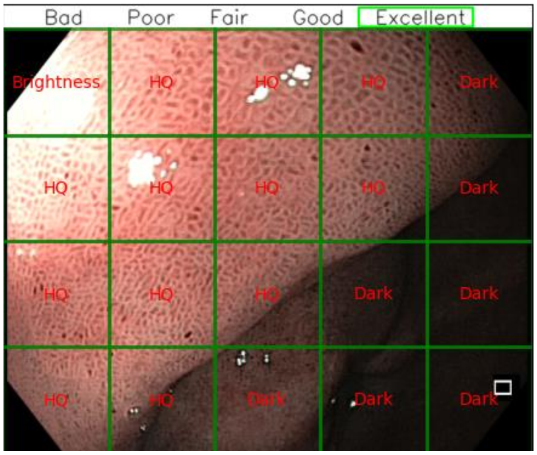


(a)

(b)

(c)

(d)



(e)

Figure 15. Descriptive results evaluate image quality: (a) Bad; (b) Poor; (c) Fair; (d) Good; (e) Excellent.

As can be seen, for images classified as bad and poor, most of the patches in the image are motion blur patches, causing almost all medical information to be lost. With respect to fair images, the majority are brightness and darkness patches, with a small number of high-quality patches remaining. As shown in Figure 15, the percentage of adjacent high-quality patches is only 25%, while that of brightness and darkness patches is 15% and 55% respectively. Finally, for good and excellent quality images, the number of motion blur patches is almost zero, and the number of high-quality patches is the majority. For instance, for good and excellent images in Figure 15, the percentage of adjacent high-quality patches is 40% and 60% respectively. Furthermore, for evaluation in the practical environment, we asked medical professionals from Viet Duc and K Tan Trieu hospitals to clinically test the results of the model. The conclusion shows that the output quality levels are highly correlated to medical professionals' perception.

5.3. Inference process:

Table 7 illustrated the time results measurement of two ways implementing inference process on that mentioned above GPU NVIDIA QUADRO RTX 4000:

Table 7. Time measurement of the original and the improved way.

Version	Time processing (mean $\pm$ std)	Frames per second (FPS)
Original	0.0848 ± 0.00149	12 FPS
Proposed	$\mathbf{0.0212 \pm 0.00138}$	**48 FPS**

By concatenating all the patches into a single tensor, we only need to perform the patch-based classification inference once instead of 20 times like the patch-by-patch classification. Therefore, by taking the refined way, the inference process of the whole pipeline increases significantly, which is four times compared to the original one.

5.4. Storage efficiency:

Table 8 illustrates the storage efficiency of the IQA module with F, G, E are the abbreviated image quality levels for Fair, Good and Excellent respectively.

Table 8. Describing the effectiveness of the IQA module in reducing storage capacity.

Video	Length	Size (mb)	Total number of frames	Quality threshold	Number of extracted frames	Amount of storage saved (%)
14-2-2018Sequence_15-14-3-228 original.avi	1m29s	123.6 mb	2244	F, G, E	950	57.66%
				G, E	625	72.15%
				E	451	79.90%
Azoulay 28032018.mp4	1m04s	40.8 mb	1623	F, G, E	298	81.64%
				G, E	237	85.40%
				E	164	89.90%

Additionally, to be able to demonstrate the effectiveness of the module in saving storage space through the amount of storage saved. The amount of storage saved is calculated by:

$$\text{The amount of storage saved} = (1 - \frac{number\ of\ extracted\ frames}{total\ number\ of\ frames}) * 100$$

where the number of extracted frames is taken according to the quality threshold. For instance, if the quality threshold is set to F, G, and E, all frames of Fair, Good, and Excellent quality will be stored, and the remaining frames of substandard quality will be discarded. Such threshold adjustment provides users a certain amount of flexibility. As can be seen from Table 8, with the first video, the amount of storage saved ranges from 57.66% to 79.90%, and the second video is from 81.64% to 89.90% depending on the quality threshold. Overall, the IQA module not only lessens the storage capacity needed but only provides users flexibility according to different scenarios.

6. Conclusion and Future works:

Given the abundance of medical imaging data nowadays, specifically in the field of gastroscopy, endoscopic image quality control plays a fundamental role in improving the overall quality of life for individuals. By cleaning and removing unsatisfactory images, image quality control not only has benefits in reducing the storage capacity needed in medical capacity but also shortening the time required in search the high-quality image, therefore helping to enhance the reviewing process and subsequent treatment. In addition, image quality assessment can be considered as supporting materials for training and evaluating young surgeons and technicians. In this work, based on the characteristics of endoscopic images, we utilized a two-stages approach to assessing the image quality using deep learning approaches. Specifically, to estimate the quality of multiple regions, first, the image is divided into non-overlap patches and utilized multi-layer features of CNN to classify into different categories, then the final quality of the entire image is determined by aggregation process, which is based on the statistical method. In addition, we also leveraged the feature magnitude to make the patch-based classification model better separating different types of patches. Lastly, we introduced an efficient inference method to improve the speed of the whole pipeline significantly. By leveraging the improved methods outlined above, we achieve nearly 98% overall accuracy for the patch classification model which is highly appreciated by medical professionals. Regarding storage, the IQA module not only decreases the required storage space by about 90% but also gives users a great deal of flexibility depending on various scenarios. In terms of processing speed, the frame rate of the pipeline reaches 47 FPS on the GPU NVIDIA QUADRO RTX4000, ensuring it runs on edge devices.

Currently, due to the fixed size of patches and non-overlap dividing strategy, it is inevitable that patches have quality ambiguity, such as there are some patches that are both high-quality and darkness. In the future, we plan to do more research on finding a more effective dividing patches strategy to handle this drawback. Additionally, because the output of the module is only evaluated by a small number of medical professionals, it may not be reliable enough to test its effectiveness in the practical environment. Therefore, we plan to consult with more experts, thereby providing a development direction for the product.

References:

[1] Petersen MW, Schroder A, Jorgensen T, et al. Irritable bowel, chronic widespread pain, chronic fatigue and related syndromes are prevalent and highly overlapping in the general population: DanFunD. Sci Rep 2020;10:3273.

[2] Spiegel B, Harris L, Lucak S, et al. Developing valid and reliable health utilities in irritable bowel syndrome: results from the IBS PROOF Cohort. Am J Gastroenterol 2009;104:1984–91.

[3] Sperber AD, Bangdiwala SI, Drossman DA, et al. Worldwide prevalence and burden of functional gastrointestinal disorders, results of Rome Foundation global study. Gastroenterology 2020.

[4] Jones MP, Crowell MD, Olden KW, Creed F. Functional gastrointestinal disorders: an update for the psychiatrist. Psychosomatics 2007;48:93–102.

[5] Shivaji UN, Ford AC. Prevalence of functional gastrointestinal disorders among consecutive new patient referrals to a gastroenterology clinic. Frontline Gastroenterol 2014;5:266–71.

[6] Aziz I, Palsson OS, Tornblom H, et al. The prevalence and impact of overlapping rome iv-diagnosed functional gastrointestinal disorders on somatization, quality of life, and healthcare utilization: a cross-sectional general population study in three countries. Am J Gastroenterol 2018;113:86–96.

[7] Mahon J, Lifschitz C, Ludwig T, et al. The costs of functional gastrointestinal disorders and related signs and symptoms in infants: a systematic literature review and cost calculation for England. BMJ Open 2017;7:e015594.

[8] Sung, Hyuna, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries, 4 Feb. 2021.

[9] Le Nga, “Medical experts identify the gut as the most vulnerable organ for Vietnamese”, VnExpress, 2019

[10] H. Inoue, N. Fukami, T. Yoshida, S.E. Kudo, Endoscopic mucosal resection for esophageal and gastric cancers, J. Gastroenterol. Hepatol. 17 (2002) 382–388.

[11] V. Catalano, R. Labianca, G.D. Beretta, G. Gatta, F. de Braud, E. Van Cutsem, Gastric cancer, Crit. Rev. Oncol. Hematol. 71 (2009) 127–164.

[12] Dinis-Ribeiro M, da Costa-Pereira A, Lopes C, LaraSantos L, Guilherme M, Moreira-Dias L, Lomba-Viana H, Ribeiro A, Santos C, Soares J, Mesquita N, Silva R, Lomba-Viana R. Magnification chromoendoscopy for the diagnosis of gastric intestinal metaplasia and dysplasia.

Gastrointest Endosc 2003.

[13] “Achieve More with NBI", Olympus, 4 May 2017

[14] Gono, Kazuhiro; Obi, Takashi; Yamaguchi, Masahiro; Ohyama, Nagasaki; Machida, Hirohisa; Sano, Yasushi; Yoshida, Shigeaki; Hamamoto, Yasuo; Endo, Takao (May 1, 2004). "Appearance of enhanced tissue features in narrow-band endoscopic imaging". Journal of Biomedical Optics. 9 (3): 568–577.

[15] Ang TL, Pittayanon R, Lau JYW, Rerknimitr R, Ho SH, Singh R, Kwek ABE, Ang DSW, Chiu PWY, Luk S, Goh KL, Ong JPL, Tan JY-L, Teo EK, Fock KM. A multicenter randomized comparison between high-definition white light endoscopy and narrow band imaging for detection of gastric lesions. Eur J Gastroenterol Hepatol 2015

[16] Gwang Ha Kim, et al. Effort to increase image quality during endoscopy: The role of pronase. National Library of Medicine (2016)

[17] Zhiyun Xue, et al. Image Quality Classification for Automated Visual Evaluation of Cervical Precancer National Library of Medicine (2022)

[18] “What happens on the day”, Gastroscopy, NHS, 2022

[19] Schoeffmann, K., Del Fabro, M., Szkaliczki, T. et al. Keyframe extraction in endoscopic video. Multimed Tools Appl 74, 11187–11206 (2015).

[20] Biorex Diagnostics, “The Importance of Diagnostic Testing”, 2023.

[21] Zhou Wang. “Applications of Objective Image Quality Assessment Methods”. IEEE Signal Processing Magazine 2011. p137-142

[22] Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4): 600–612.

[23] Sewoong Ahn, Yeji Choi, and Kwangjin Yoon. Deep learning-based distortion sensitivity prediction for full reference image quality assessment. In IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 344–353, 2021.

[24] Keyan Ding, Kede Ma, Shiqi Wang, and Eero Simoncelli. Image quality assessment: Unifying structure and texture similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

[25] Seyed Ayyoubzadeh and Ali Royat. (ASNA) An attention based siamese-difference neural network with surrogate ranking loss function for perceptual image quality assessment. In IEEE

Conference on Computer Vision and Pattern Recognition Workshops, pages 388–397, 2021

[26] S Alireza Golestaneh, Saba Dadsetan, and Kris M Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022.

[27] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Proc. of NeurIPS, 2017.

[29] Junyong You and Jari Korhonen. Transformer for image quality assessment. In Proc. of ICIP, 2021

[30] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In Proc. of ICCV, 2021.

[31] Wang, Z.: Objective image quality assessment: facing the real-world challenges. In: IS&T International Symposium on Electronic Imaging, pp. 1–6 (2016).

[32] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, “End-to-end blind image quality assessment using deep neural networks,” IEEE Transactions on Image Processing, vol. 27, no. 3, pp. 1202–1213, 2017.

[33] K.-Y. Lin and G. Wang, “Hallucinated-IQA: No-reference image quality assessment via adversarial learning,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 732–741, 2018.

[34] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, “MetaIQA: deep meta-learning for no-reference image quality assessment,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.

[35] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, “Blindly assess image quality in the wild guided by a self-adaptive hyper network,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3667– 3676, 2020.

[36] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In Proc. of CVPR, 2020.

[37] Yunhong Li, Huanhuan Zhang, Jinni Chen, Peng Song, Jie Ren, QiuMing Zhang, KaiLi Jia. "Non-reference image quality assessment based on deep clustering". Signal Processing: Image Communication, Volume 83, 2020.

[38] Kong, X., Yang, Q. No-Reference Image Quality Assessment for Image Auto-Denoising. Int J Comput Vis 126, 537–549 (2018).

[39] Xialei Liu, Joost van de Weijer, Andrew D. Bagdanov. "RankIQA: Learning from Rankings for No-reference Image Quality Assessment". Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1040-1049

[40] Sato, T. "Texture and Color Enhancement Imaging for Endoscopic Screening". Journal of Healthcare Engineering, 2021, 5518948 (2021).

[41] Eric W Abel, Nikolaos Fotiadis, Paul S White. "Light intensity distribution in images from rigid endoscopes used in minimal access sinus surgery", Laryngoscope Investig Otolaryngol, 2021 Nov 23;6(6):1283-1288.

[42] Tan W, Xu C, Lei F, Fang Q, An Z, Wang D, Han J, Qian K, Feng B. An Endoscope Image Enhancement Algorithm Based on Image Decomposition. *Electronics*. 2022; 11(12):1909.

[43] Yang, X., Chen, Y., Tao, R., Zhang, Y., Liu, Z., & Shi, Y. (2020). Endoscopic Image Deblurring and Super-Resolution Reconstruction Based on Deep Learning. 2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE).

[44] Fei Gao, Yi Wang, Panpeng Li, Min Tan, Jun Yu, and Yani Zhu. Deepsim: Deep similarity for image quality assessment. Neurocomputing, 257:104–114, 2017

[45] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In CVPR, 2018.

[46] Sebastian Bosse, Dominique Maniry, Klaus-Robert Muller, ¨ Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. IEEE TIP, 2017.

[47] Li, Y.; Liu, L. Image quality classification algorithm based on InceptionV3 and SVM. MATEC Web Conf. 2019, 277, 02036.

[48] Golchubian, A., Marques, O. & Nojoumian, M. Photo quality classification using deep learning. Multimed Tools Appl 80, 22193–22208 (2021).

[49] Feng Li Yu, Jing Sun, Annan Li, Jun Cheng, Cheng Wan, and Jiang Liu, "Image quality classification for DR screening using deep learning," Conf. Proc. IEEE Eng. Med. Biol. Soc.,

vol. 2017, pp. 664–667, Jul. 2017.

[50] Xue Z, Angara S, Guo P, Rajaraman S, Jeronimo J, Rodriguez AC, Alfaro K, Charoenkwan K, Mungo C, Domgue JF, Wentzensen N, Desai KT, Ajenifuja KO, Wikström E, Befano B, de Sanjosé S, Schiffman M, Antani S. Image Quality Classification for Automated Visual Evaluation of Cervical Precancer. Milland Workshop (2022). 2022 Sep; 13559:206-217.

[51] F. Yu, J. Sun, A. Li, J. Cheng, C. Wan and J. Liu, "Image quality classification for DR screening using deep learning," 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jeju, Korea (South), 2017, pp. 664-667

[52] Mahapatra, Dwarikanath, et al. "Retinal image quality classification using saliency maps and CNNs." International Workshop on Machine Learning in Medical Imaging. Cham: Springer International Publishing, 2016.

[53] Wang Y., Song Y., Wang F., Sun J., Gao X., Han Z., Shi L., Shao G., Fan M., Yang G. A two-step automated quality assessment for liver MR images based on convolutional neural network. Eur. J. Radiol. 2020; 124:108822. doi: 10.1016/j.ejrad.2020.108822

[54] Schoeffmann, K., Del Fabro, M., Szkaliczki, T. et al. Keyframe extraction in endoscopic video. Multimed Tools Appl 74, 11187–11206 (2015).

[55] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans and G. Carneiro, "Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 4955-4966, doi: 10.1109/ICCV48922.2021.00493.

[56] Taye, M.M. Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, Future Directions. Computation 2023, 11, 52.

[57] Ballester, P., & deAraújo, R. M. (2016, February)"On the Performance of GoogLeNet and AlexNet Applied to Sketches." in AAAI.

[58] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition. 2015.

[59] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Advances in Neural Information Processing Systems 32 (pp. 8024–8035).

[60] Kingma, D. and Ba, J. (2015) Adam: A Method for Stochastic Optimization. Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)