Article

Not peer-reviewed version

Video Analysis of Small Bowel Capsule Endoscopy Using a Transformer Network

SangYup Oh , DongJun Oh , Dongmin Kim , Woohyuk Song , Youngbae Hwang , Namik Cho * , Yun Jeong Lim *

Posted Date: 4 August 2023

doi: 10.20944/preprints202308.0447.v1

Keywords: artificial intelligence; Transformer; capsule endoscopy; video-analysis

Article

Video Analysis of Small Bowel Capsule Endoscopy Using a Transformer Network

SangYup Oh [1], DongJun Oh [2], Woohyuk Song [3], Youngbae Hwang [4], Dongmin Kim [5], Namik Cho [1,*] and Yun Jeong Lim [2,*]

[1] School of Electrical and Computer Engineering, Seoul National University, 1 Gwanak-ro, Kwanak-Gu, Seoul, KOREA; syup5@snu.ac.kr
[2] Department of Internal Medicine, Dongguk University Ilsan Hospital, Dongguk University College of Medicine, Goyang, Korea; mileo31@naver.com
[3] School of Electrical and Computer Engineering, Seoul National University, 1 Gwanak-ro, Kwanak-Gu, Seoul, KOREA; whsong@snu.ac.kr
[4] Department of Electronics Engineering, Chungbuk National University, Cheongju 28644, Korea; ybhwang@chungbuk.ac.kr
[5] JLK TOWER, Gangnam-gu, Seoul, 06141; dmkim@jlkgroup.com
* Correspondence: nicho@snu.ac.kr; Tel.: +82-2-880-8480; drlimyj@gmail.com; Tel.: +82-31-961-7133

Abstract: Although wireless capsule endoscopy (WCE) detects small bowel diseases effectively, it has some limitations. For example, the reading process can be time-consuming due to the numerous images generated per case, and lesion detection accuracy may rely on the operators' skills and experiences. Hence, many researchers have recently developed deep learning-based methods to address these limitations. However, they tend to select only a portion of the images from a given WCE video and analyze each image individually. In this study, we note that more information can be extracted from the unused frames and temporal relations of sequential frames. Specifically, to increase the accuracy of lesion detection without depending on experts' frame selection skills, we suggest using whole video frames as the input to the deep-learning system. Thus, we propose a new Transformer-based neural encoder that takes the entire video as the input, exploiting the power of the Transformer to extract long-term global correlation within and between the input frames. Subsequently, we can capture the temporal context of the input frames and the attentional features within a frame. Tests on benchmark datasets of four WCE videos showed 95.1% sensitivity and 83.4% sensitivity. These results may significantly advance automated lesion detection techniques for WCE images. Our code is available at https://github.com/syupoh/VWCE-Net.git.

Keywords: artificial intelligence; Transformer; capsule endoscopy; video-analysis

1. Introduction

With the introduction of wireless capsule endoscopy (WCE) for the first time in 2000, there has been a paradigm shift in small bowel examination [1,2]. After the patient swallows the capsule, it moves along the digestive tract from the mouth to the anus without user manipulation. Indeed, WCE is essential for monitoring Crohn's disease, small bowel bleeding, small bowel tumors, and inherited polyp syndrome, and it also has the advantage of being noninvasive and convenient for patients compared to wired endoscopy [3,4]. However, due to the location and length of the small bowel, WCE captures images for more than eight hours, generating more than 50,000 images for each examination.

Thus, even an experienced gastroenterologist would need more than two hours to review all of the small bowel images and detect any lesions [5]. This long evaluation time is the main limitation of traditional methods, resulting in high fatigue in physicians. To overcome this limitation and exploit recently developed deep-learning methods for automated lesion detection, convolutional neural

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2023 doi:10.20944/preprints202308.0447.v1

network (CNN)-based models have been proposed, which have been shown to effectively detect lesions such as bleeding, inflammation, vascular, and polyps among WCE images [6–9].

Although CNN-based image-reading models have demonstrated high accuracy in detecting lesions for a given frame, they have limitations in that only one still-shot image is used to detect lesions; thus, they do not benefit from sequential images containing more information. As shown in Figure 1, sequential images offer temporal continuity, which can provide additional information about the lesions. Furthermore, these images possess redundancy, which can help in avoiding false negatives, unlike still-shot images. In the real world, WCE images are stored and displayed as videos, where each frame is captured at an appropriate rate [10]. In addition, if a lesion is suspected during reading, the gastroenterologist not only judges the still-shot image but also refers to its adjacent frames together. Therefore, deep-learning models that can consider whole frames of WCE videos are required.

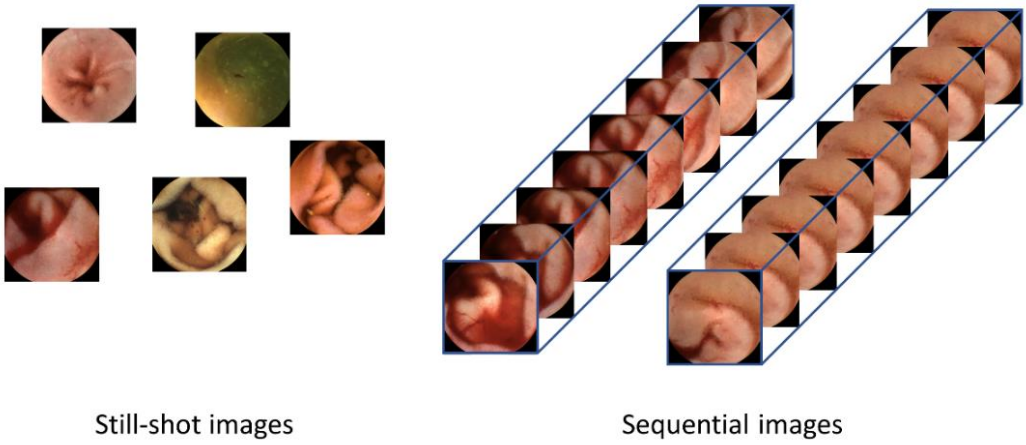

Figure 1. Difference between human-selected still-shot images and sequential images as the inputs to the lesion detection systems. Still-shot images do not have continuity between images, so temporal information cannot be used. Using the whole frames of the video has advantages in that more information can be used, and the temporal changes within the frames can provide valuable cues for detecting lesions.

Although CNNs are sufficiently effective at extracting features within a region of an image, they have limited receptive fields such that their attention is limited to a local area. On the other hand, the recently developed Transformer architecture is known to have advantages over CNNs, particularly in that the Transformer is designed to extract long-term global attention [11]. The Transformer was initially developed for natural language modeling, which achieved significant performance gains over the preceding neural networks. Furthermore, the Vision Transformer (ViT), a model that modified the original Transformer for computer vision, has also performed well in image classification [12]. Because ViT performs well in computer vision tasks, some studies have employed the Transformer architecture to analyze WCE images [13–15].

The characteristic of continuous WCE images shares significant similarities with Natural Language Processing (NLP) tasks compared to single-image processing tasks. WCE images are not in a single form; instead, they are in a sequential type with continuity. Continuous data have contextual associations. In NLP task, the meanings of words are understood differently by other words in the sentence. The recognition process of words is interdependent. In other words, words affect each other. Recurrent neural network (RNN) or Long-Short Term Memory (LSTM) architectures are designed to solve problems caused by long-range dependencies by modeling sequential data [16,17]. Similarly, Images in video have the same characteristics as words in NLP. In video understanding tasks, other images influence the understanding of one image. For example, if an object that was difficult to distinguish because it was blurry in a specific image is clearly visible in another image and can be distinguished, it becomes possible to accurately distinguish the object even

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2023 doi:10.20944/preprints202308.0447.v1

though there is an image in which the object appeared blurry. For this reason, there have been attempts in video task to learn by making temporal information into features, such as 3D convolution [18–21]. Furthermore, since WCE videos contain relatively many low-quality images compared to other videos, the impact of other images on understanding images is more important. Therefore, it can be expected that the Transformer-based long-range self-attention model would be more effective for WCE video analysis [22].

In this paper, we propose a novel video Transformer model that utilizes the self-attention of images and video information to detect small bowel disease. Unlike the traditional image-level analysis that relies on human-selected images as the input, we conduct a video-level analysis using our proposed Video Wireless Capsule Endoscopic Network (VWCE-Net). Specifically, our method performs video-level analysis by considering all frames of the video as the input. This model performs video analysis based on spatio-temporal self-attention information obtained from sequential images rather than one still-shot image, as in conventional methods. In the experiments, we compare the proposed method with well-known image-based models such as the YOLOv4 and Xception models [23,24].

2. Materials and Methods

2.1. Data Acquisition (Data Characteristics)

A total of 260 WCE (MiroCam MC1600, Intromedic Co., Ltd., Seoul, Korea) cases performed at Dongguk University Ilsan Hospital between 2002 and 2022 are used to train the proposed Transformer model. All WCE image frames are stored in JPEG format with dimensions of 320 × 320 and a frame rate of 3 fps using MiroView 4.0 (Intromedic Co., Ltd., Seoul, Korea). Our study was conducted with the approval of the Institutional Review Board of Dongguk University Ilsan Hospital (DUIH 2018-10-009-010).

2.2. Data Preparation

Two gastroenterologists specializing in capsule endoscopy (Oh DJ and Lim YJ from Dongguk University Ilsan Hospital) independently performed image labeling for lesion detection. After manually reviewing and categorizing the entire WCE case images as normal, bleeding, inflammation, vascular, and polyp tissues, they cross-checked their findings to ensure accuracy. The 40 labeled case datasets are then classified into training (36 cases, 90%) and test (4 cases, 10%) sets. The datasets were composed of clip units rather than still-shot images. Each clip consisted of four sequential images. The reason for setting the unit of one clip as four sequential images is discussed in Section 3.2. The training set consists of 1,291,004 (322,751), 140,788 (35,197), 10,912 (2,728), 2,328 (582), and 14,832 (3,708) images (clips) for normal, bleeding, inflammation, vascular, and polyp tissues, respectively. The test set consists of 172,820 (43,205), 4200 (1,050), 304 (76), 892 (223), and 24 (6) images (clips) for normal, bleeding, inflammation, vascular, and polyp tissues, respectively (Table 1).

Table 1. Dataset specifications.

	Whole Dataset		Training Dataset		Test Dataset	
Classes	**Clips**	**Images**	**Clips**	**Images**	**Clips**	**Images**
Normal	365,956	1,463,824	322,751	1,291,004	43,205	172,820
Bleeding	36,247	144,988	35,197	140,788	1,050	4,200
Inflammation	2,804	11,216	2,728	10,912	76	304
Vascular	805	3,220	582	2,328	223	892
Polyp	3,714	14,856	3,708	14,832	6	24
Cases		40		36		4

In previous studies, labeling was traditionally conducted on an individual basis to identify the presence of lesions. However, in this study, we aim to reduce manual labor costs by implementing a simultaneous labeling approach for sequential images, as depicted in Figure 2. Nonetheless, utilizing

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2023 doi:10.20944/preprints202308.0447.v1

sequential images as a dataset presents certain limitations. Firstly, the presence of numerous similar images within a video sequence poses challenges to effective learning. Learning from a dataset containing numerous similar images increases the risk of overfitting, which can hinder the model's generalization capability. Secondly, there is a significant data imbalance between normal and abnormal images. The dataset primarily comprises normal images, while the number of abnormal images is relatively low. This data imbalance can negatively impact the performance of the model, leading to suboptimal results.

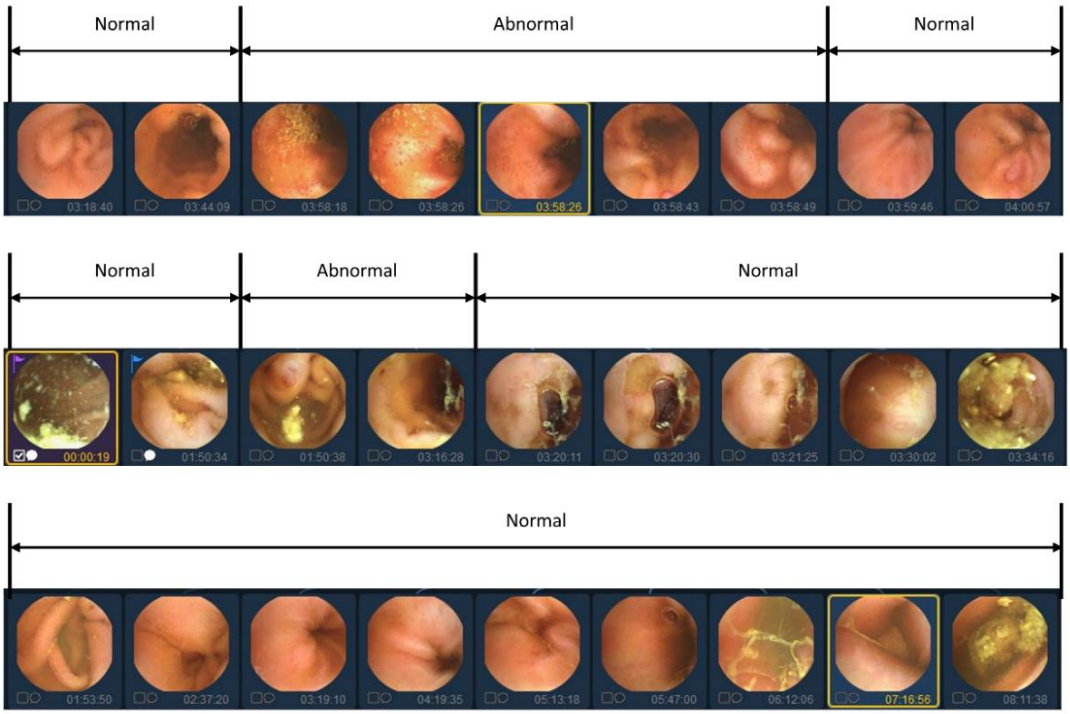

Figure 2. Video data labeling form with abnormal segment display.

2.3. Study Design

Our method is implemented using the PyTorch codebase MMAction2 [25]. Figure 2 shows a flow chart of the proposed VWCE-Net. First, the sequential input images are converted into clips comprising several images. Next, the clips are placed into the model, and the lesion is detected in each clip.

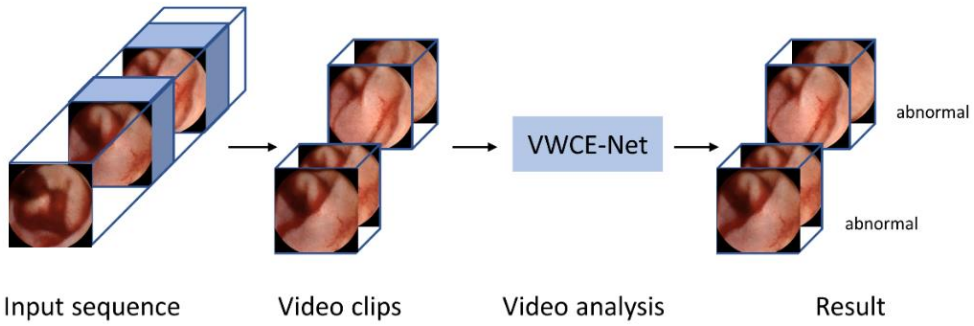

Figure 3. Flowchart of the proposed VWCE-Net.

In this study, we set the clip size to 4 (that is, four consecutive frames constituted a clip). The size of each frame is 320 pixels in width and height, but we resize them to 224 pixels. Unlike ViT [12], which takes a single image as input, our method's input is a clip containing multiple images. Hence, the dimension of the input is also different from conventional methods, precisely 4 × 224 × 224 × 3

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2023 doi:10.20944/preprints202308.0447.v1

[22], where 4 represents the number of images in the clip. Specifically, the input sequence of VWCE-Net is $x \in R^{4\times224\times224\times3}$, and as shown in Figure 4., one 224 × 224 image becomes 14×14 $(N = 196)$ patches, denoted as $x_p \in R^{4\times196\times16\times16\times3}$ with a patch size of 16×16. These patches are flattened into $x_{fp} \in R^{4\times196\times768}$.

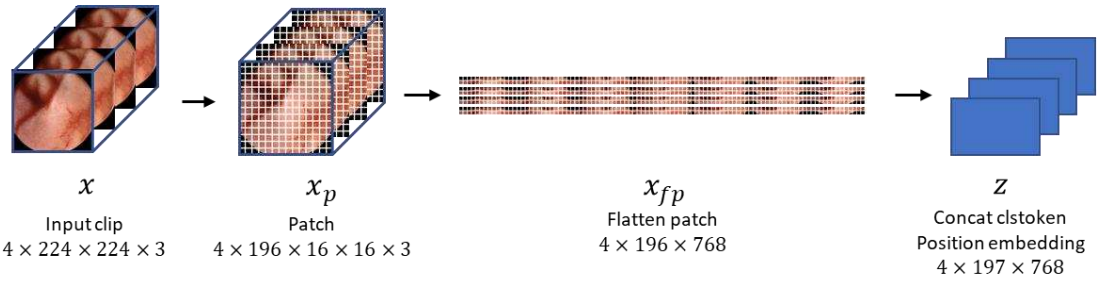

Figure 4. The process by which an input clip becomes an embedded feature.

In BERT, clstoken is added to first of all features, used in classification tasks, and ignored in other tasks [26]. After passing through all layers of the Transformer, the "clstoken" acquires the combined meaning of the token sequence. In the classification task, you can pass through this clstoken to the classifier to classify the entire sentence entered. In contrast to BERT, where the input data was in word form, in this work, the input embedding is of the dimension $x_{fp} \in R^{4\times196\times768}$, so the clstoken is represented as $t_{cls} \in R^{4\times1\times768}$. The clstoken t_{cls} is concatenated to the flattened patch x_{fp}. Finally, the size of the embedding becomes $R^{4\times(196+1)\times768}$. VWCE-Net learns this clstoken to represent a sequence composed of 196 patched images so that it can operate as a classification token that determines whether there is a lesion or not.

The Transfomer-based self-attention model does not compute convolution and does not contain recurrence like LSTM. To use the order of sequence information, it is necessary to mathematically model the relative or absolute position of flatten patches.

$$v_{pos_{(t,i)}} \begin{cases} \sin(t \times w_k), if\ i = 2k \\ \cos(t \times w_k), if\ i = 2k+1 \end{cases}$$
$$w_k = \frac{1}{10000^{2k/d}}\ (d\ is\ dimension\ of\ embedding) \tag{1}$$

Where t is the position in the sequence and i is the index of the dimension representing the position. Positional encoding v_{pos} has the form of sinusoidal. It has a pair value of sine and cosine depending on the value of i. That is, the even-numbered dimension uses sin and the odd-numbered dimension uses cos.

$$v_{pos} = \left[\begin{pmatrix} v_{pos_{(0,0)}} & v_{pos_{(0,1)}} & \cdots & v_{pos_{(0,767)}} \\ \vdots & \ddots & \ddots & \vdots \\ v_{pos_{(196,0)}} & v_{pos_{(196,1)}} & \cdots & v_{pos_{(196,767)}} \end{pmatrix}\right]$$
$$= \left[\begin{pmatrix} \sin(0) & \cos(0) & \cdots & \sin(0/10000) \\ \sin(1) & \cos(1) & \cdots & \sin(1/10000) \\ \vdots & \ddots & \ddots & \vdots \\ \sin(196) & \cos(196) & \cdots & \sin(196/10000) \end{pmatrix}\right] \tag{2}$$

By this poitional encoding, in NLP, even the same word can have different embedding values depending on the position used in the sentence [11]. In tasks such as text translate and text generation, 1-dimensional positional encoding is calculated because the input is a 1-dimensional word. In the task of this experiment to find a lesion in a given image, since the input is a 2-dimensional image, 2-dimensional position encoding can be considered. To apply 2-dimensional position encoding, first divide embedding in half, set one to X-embedding and the other to Y-embedding. Each size is set to $d/2$. By concatenating X-embedding and Y-embedding, the final positional encoding value of the patch of the corresponding position can be obtained. This work uses 1-dimensional positional encoding instead of 2-dimensional positional encoding because there is little difference in performance [12]. This means that 2-dimensional positional relationship information between

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2023 doi:10.20944/preprints202308.0447.v1

coordinate of X and Y is sufficiently included in the 1-dimensional positional relationship between flatten patches.

Then, the embedded feature $z \in R^{4\times197\times768}$ is created by adding the position embedding vector $v_{pos} \in R^{4\times197\times768}$, which includes the spatial information of the patch. Formally, the embedded feature z is represented as

$$z = [t_{cls}||x_{fp}] + v_{pos} \tag{3}$$

where $||$ represents the concatenation.

The embedded feature $z \in R^{4\times197\times768}$ is projected as a query q, key k, and value v representations following the Transformer architecture [11]. This transformation is achieved through linear operations using parameter matrices $W_{qkv} \in R^{768\times(768\times3)}$, which are described as

$$q, k, v = zW_{qkv} \tag{4}$$

where the dimensions of the projected $q, k, and\ v$ are equally $4 \times 197 \times 96$. Each operation of $q, k, and\ v$ contains layer normalization of embedded feature z.

As shown in Figure 5, the Multi-Head Attention module performs several self-attention operations in parallel. Each of these operations is referred to as a "Head," following the terminologies in [11]. In this paper, the number of "Heads," denoted as $\mathcal{A}$, is set to 8. Because the feature dimension is 768, the dimension within the multi-head is 96. Consequently, the $q, k,$ and v vectors are converted to dimensions of $4 \times 8 \times 197 \times 96$.

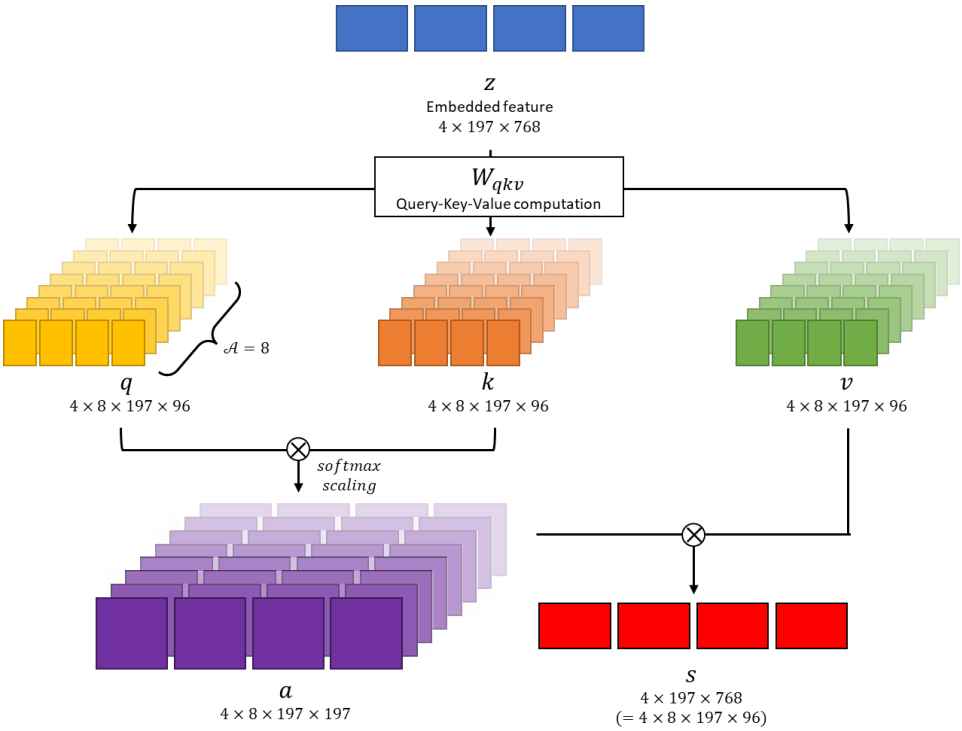

Figure 5. Multi-Head Attention module overview.

We perform matrix multiplication between q and k, followed by scaling with D_h, where $D_h = D/\mathcal{A}$, and then apply the softmax function as [26]

$$a = softmax\left(\frac{q \otimes k^T}{\sqrt{D_h}}\right) \tag{5}$$

which is the self-attention coefficient. In this paper, we set $D_h = 96$. Then, the self-attention value is obtained as

$$s = a \otimes v \tag{6}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2023 doi:10.20944/preprints202308.0447.v1

$\otimes$ denotes element-wise product. Multi-head attent heads are concatenated and pass through the Multi Layer Perceptron (MLP). For each operation residual connection is used.

$$z' = [s_1||s_2||\cdots||s_\mathcal{A}] + z$$
$$z^1 = \mathrm{MLP}(\mathrm{LN}(z')) + z' \quad (7)$$

LN denotes layer normalization and MLP is a multi-layer perceptron consisting of two hidden layers. In summary, clstoken t_{cls} is concatenated to the flatten image patch x_{fp} and positional encoding v_{pos} is added to obtain embedded feature z. Then, z is used to calculate multi head attention to obtain self attention s, and s passes through MLP to obtion z^1 in the Transformer layer. In this paper, we set the number of Transformer layers to 12, so we calculate up to z^{11} by repeating Transformer operation. The first position of z^{11} obtained through the entire Transformer layer is used to determine whether there is a lesion or not.

$$Y = FC(LN(z_{(0)}^{11})) \quad (8)$$

FC denotes fully connected layer $\in R^{768\to5}$. If the classification result is 0, it means that there is no lesion, and if it is more than 1, it means that there is a lesion. The reason why the final output dimension is 5 is because the type of lesion is set to 4 when preparing the data.

2.4. Implementation Detail

The basic network that makes the images into patches and passes them through the transformer network is the same as the base ViT architecture [12]. Base ViT was pretrained with ImageNet-21K, and in this experiment, 1,291,004 images were additionally trained and fine-tuned. During training, the train images are flipped, and the flip ratio is set to 0.5. The momentum is set to 0.9 and the weight decay is set to 0.0001. The learning rate is set to 0.005 and the maximum epoch was set to 100.

3. Results

3.1. Comparison of Our Video-Level Analysis with Existing Still-Shot Image Analysis Methods

The results of the proposed video classification, image classification, and object detection models are compared in Table 2.. The model used for classification was XceptionNet [27], and the model used for object detection was YOLOv4 [28]. Existing methods for classification and object detection rely on image-level analysis [23,24], which involves only human-selected images from the entire video. In contrast, we evaluate the entire video in this experiment instead of relying on a human-selected subset.

The sensitivity and specificity performances of the three models are compared with the same full video test dataset. Table 2. shows that the classification model XceptionNet has a low sensitivity rate, and the object detection model YOLOv4 has a low specificity rate compared to the proposed model VWCE-Net.

Table 2. Comparative results of the video analysis and image-based methods.

Model	Sensitivity	Specificity
VWCE-Net	95.1	83.4
XceptionNet	43.2	90.4
YOLOV4	88.6	51.5

3.2. Determining the Clip Length for the Video-Level Analyses

To investigate the impact of clip length in our video-level analysis, we conduct experiments by dividing the clip length by 2, 4, 6, and 8. The results of these comparisons are presented in Table 3.. Surprisingly, we can see that the sensitivity and specificity are not directly proportional to the clip length. Consequently, it is crucial to determine whether to prioritize sensitivity or specificity in determining the optimal clip size. After careful evaluation, we determine the clip size as 4, which

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2023 doi:10.20944/preprints202308.0447.v1

yields the highest specificity while maintaining a sufficiently high sensitivity. Therefore, all the explanations provided earlier and the subsequent experiments are conducted based on a clip size of 4.

Table 3. Comparative results based on clip length.

Clip length	Sensitivity	Specificity
2	98.51	72.81
4	95.13	83.43
6	93.93	83.25
8	97.89	80.76

3.3. Sampling Strategy

To analyze the WCE images more accurately, sequential images are used as the dataset instead of still-shot images. However, there are limitations in using sequential images as a dataset. Firstly, sequential images often exhibit similarities, particularly between adjacent frames. The abundance of similar images poses a challenge to effective learning and increases the risk of overfitting. Secondly, the dataset suffers from an imbalance, with a substantial number of normal images compared to abnormal ones. This imbalance negatively impacts the model's performance. To address these challenges, a sampling strategy is employed to select only a portion of the WCE images as clips for training.

When studying a model for action recognition using video, it is common practice to utilize only the clip in the center [29]. However, in this study, in addition to the center sampling of the clip, random sampling is also performed to determine if there is a difference depending on the sampling method. Figure 6 illustrates the scenarios of center sampling and random sampling. A comparison of the random sampling and center sampling approaches is shown in Table 4, which clearly indicates that center sampling yields superior performance.

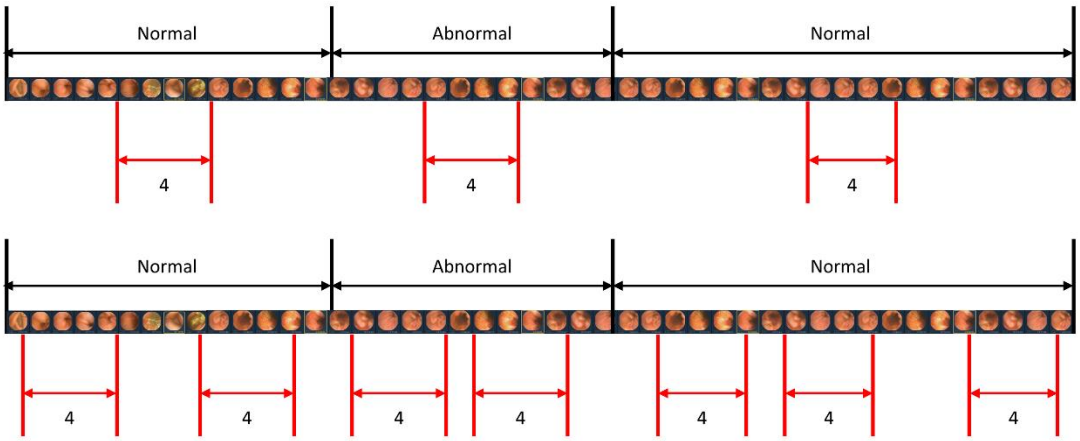

Figure 6. Sampling methods for video input. The upper image represents the center sampling method, where only the center part of the video sequence is extracted as a clip. The lower image demonstrates the random sampling method, where clips are randomly extracted from the entire video sequence.

Table 4. Performance comparison: Center sampling vs. random sampling (clip length: 4).

Sampling Strategy	Sensitivity	Specificity
center	95.13	83.43
random	72.84	89.62

 doi:10.20944/preprints202308.0447.v1

4. Discussion

4.1. The Necessity of Utilizing Transformer Networks and Video Input

This study is the first to analyze complete video frames from capsule endoscopy using a Transformer network. This approach is motivated by the observation that gastroenterologists typically identify diseases not solely based on individual images but by considering a sequence of images surrounding a suspected area [30]. However, in existing studies that automatically detect symptoms using AI models, the inputs to the system were usually single images as opposed to several consecutive images [23,31,32]. Therefore, it is necessary to manipulate WCE images in a continuous manner to set up an environment similar to actual disease detection.

The Transformer is currently a state-of-the-art model for Computer Vision and NLP tasks. As a result, recent works [13,15] have introduced the application of Transformers for processing and analyzing WCE images. These studies have demonstrated the effectiveness of utilizing the Transformer architecture in achieving high performance when applied to WCE images.

The results of the proposed video classification, image classification, and object detection models are compared in Table 2.. The model used for classification was XceptionNet [27], and the model used for object detection was YOLOv4 [28]. Notably, existing methods for classification and object detection rely on image-level analysis [23,24], which involves only human-selected images from the entire video. In contrast, we evaluate the entire video in this experiment instead of relying on a human-selected subset. Hence, the sensitivity and specificity performances of the three models are compared using the same full video test dataset. As a result, it is shown that the classification model XceptionNet exhibits a low sensitivity rate, while the object detection model YOLOv4 exhibits a low specificity rate compared to the proposed VWCE-Net model (Table 2). Moreover, Table 2 provides a comprehensive comparison of the detection results between our proposed VWCE-Net and the image-based CNN approach. It can be seen that our video-based method significantly outperforms the image-based methods in terms of accuracy.

4.2. Model Setup Analysis

To accurately detect lesions within a clip, we obtain detection results by changing the length of the clip. Additionally, the length of the clip is modified for each result to ensure optimal detection. This is because when the clip is too long, the number of images with lesions in the clip exceeds those without lesions. Hence, to achieve an exemplary detection result, it is crucial to ensure an appropriate clip length that maintains continuity without being too long. As demonstrated in Table 3, it is evident that a clip length of 4 yields the best results. Accordingly, four images are concatenated into one clip to determine whether there is a lesion in each clip. In addition, considering the considerable length of capsule endoscopy videos, we sample only a portion of the images from the videos for training. Two sampling strategies, center sampling and random sampling, are employed in this process. As illustrated in Figure 6, the video is segmented into sequences whenever the presence or absence of a lesion changes. The center sampling method extracts only the central images from each sequence to form a clip, while the random sampling method involves the random extraction of images, irrespective of their position within the sequence.

The center sampling method is widely utilized in video recognition tasks, as it assumes that the images positioned at the center of the sequence are representative of the entire sequence. Conversely, the images extracted through random sampling lack the same level of representativeness. Consequently, incorporating randomly extracted images does not significantly improve the model's performance. As shown in Table 4, the model trained with randomly extracted images has a lower false positive rate than the model trained with center-extracted images. However, training with center-extracted images yields higher sensitivity.

While randomly extracted images introduce redundant information that restricts the model's performance improvement, relying solely on center-extracted images is not considered optimal due to the limited information they provide. Therefore, striking a balance between the two sampling methods is crucial to achieving better detection outcomes.

5. Conclusions

In this paper, we have proposed a Transformer-based neural network for lesion detection in WCE data. Unlike conventional methods that analyze still-shot images, our approach analyzes sets of consecutive images. We have trained and tested our network using a dataset of approximately 1.6 million WCE images, which were categorized into five classes: normal, bleeding, inflammation, vascular, and polyp. To evaluate the performance of our model, we compared it with existing image-based methods, specifically the XceptionNet and YOLOv4 models. The experimental results clearly demonstrate that our video-based VWCE-Net model outperforms the image-based methods in terms of lesion detection accuracy and effectiveness. By leveraging the power of Transformer-based architecture and analyzing WCE data using consecutive image sequences, our proposed approach offers improved performance and holds promise for advancing lesion detection in WCE examinations.

Author Contributions: For research articles with several authors, the following statements should be used "Conceptualization, X.X. and Y.Y.; methodology, X.X.; software, X.X.; validation, X.X., Y.Y. and Z.Z.; formal analysis, X.X.; investigation, X.X.; re-sources, X.X.; data curation, X.X.; writing—original draft preparation, X.X.; writing—review and editing, X.X.; visu-alization, X.X.; supervision, X.X.; project administration, X.X.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.".

Funding: This study was supported by a grant from the Korean Health Technology R&D Project through the Korean Health Industry Development Institute (KHIDI, https://www.khidi.or.kr/eps), funded by the Ministry of Health & Welfare, Republic of Korea (Grant Number: HI19C0665).

Institutional Review Board Statement: In this section, you should add the Institutional Review Board Statement and approval number, if relevant to your study. You might choose to exclude this statement if the study did not require ethical approval. Please note that the Editorial Office might ask you for further information. Please add "The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval)." for studies involving humans. OR "The animal study proto-col was approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval)." for studies involving animals. OR "Ethical review and approval were waived for this study due to REASON (please provide a detailed justification)." OR "Not applicable" for studies not involving hu-mans or animals.

Informed Consent Statement: Any research article describing a study involving humans should contain this statement. Please add "Informed con-sent was obtained from all subjects involved in the study." OR "Patient consent was waived due to REASON (please provide a detailed justification)." OR "Not applicable." for studies not involving humans. You might also choose to exclude this statement if the study did not involve humans. Written informed consent for publication must be obtained from participating patients who can be identified (in-cluding by the patients themselves). Please state "Written informed consent has been obtained from the patient(s) to publish this paper" if applicable.

Data Availability Statement: We encourage all authors of articles published in MDPI journals to share their research data. In this section, please provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study. Where no new data were created, or where data is unavailable due to privacy or ethical restrictions, a statement is still required. Suggested Data Availability Statements are avail-able in section "MDPI Research Data Policies" at https://www.mdpi.com/ethics.

Conflicts of Interest: Declare conflicts of interest or state "The authors declare no conflict of interest.".

References

1. Soffer, S.; Klang, E.; Shimon, O.; Nachmias, N.; Eliakim, R.; Ben-Horin, S.; Kopylov, U.; Barash, Y. Deep learning for wireless capsule endoscopy: a systematic review and meta-analysis. *Gastrointest Endosc* **2020**, *92*, 831-839 e838, doi:10.1016/j.gie.2020.04.039.
2. Iddan. Wireless capsule endoscopy. *Nature* **2000**.
3. Eliakim, R. Video capsule endoscopy of the small bowel. *Curr Opin Gastroenterol* **2010**, *26*, 129-133, doi:10.1097/MOG.0b013e328334df17.
4. Pennazio, M.; Spada, C.; Eliakim, R.; Keuchel, M.; May, A.; Mulder, C.J.; Rondonotti, E.; Adler, S.N.; Albert, J.; Baltes, P.; et al. Small-bowel capsule endoscopy and device-assisted enteroscopy for diagnosis and

treatment of small-bowel disorders: European Society of Gastrointestinal Endoscopy (ESGE) Clinical Guideline. *Endoscopy* **2015**, *47*, 352-376, doi:10.1055/s-0034-1391855.
5. Committee, A.T.; Wang, A.; Banerjee, S.; Barth, B.A.; Bhat, Y.M.; Chauhan, S.; Gottlieb, K.T.; Konda, V.; Maple, J.T.; Murad, F.; et al. Wireless capsule endoscopy. *Gastrointest Endosc* **2013**, *78*, 805-815, doi:10.1016/j.gie.2013.06.026.
6. Jia, X.; Xing, X.; Yuan, Y.; Xing, L.; Meng, M.Q.H. Wireless Capsule Endoscopy: A New Tool for Cancer Screening in the Colon With Deep-Learning-Based Polyp Recognition. *Proceedings of the IEEE* **2020**, *108*, 178-197, doi:10.1109/jproc.2019.2950506.
7. Kim, S.H.; Hwang, Y.; Oh, D.J.; Nam, J.H.; Kim, K.B.; Park, J.; Song, H.J.; Lim, Y.J. Efficacy of a comprehensive binary classification model using a deep convolutional neural network for wireless capsule endoscopy. *Scientific Reports* **2021**, *11*, 17479.
8. Kim, S.H.; Lim, Y.J. Artificial intelligence in capsule endoscopy: A practical guide to its past and future challenges. *Diagnostics* **2021**, *11*, 1722.
9. Oh, D.J.; Hwang, Y.; Lim, Y.J. A Current and Newly Proposed Artificial Intelligence Algorithm for Reading Small Bowel Capsule Endoscopy. *Diagnostics (Basel)* **2021**, *11*, doi:10.3390/diagnostics11071183.
10. Spada, C.; McNamara, D.; Despott, E.J.; Adler, S.; Cash, B.D.; Fernández-Urién, I.; Ivekovic, H.; Keuchel, M.; McAlindon, M.; Saurin, J.-C. Performance measures for small-bowel endoscopy: a European Society of Gastrointestinal Endoscopy (ESGE) quality improvement initiative. *Endoscopy* **2019**, *51*, 574-598.
11. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
12. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**.
13. Bai, L.; Wang, L.; Chen, T.; Zhao, Y.; Ren, H. Transformer-Based Disease Identification for Small-Scale Imbalanced Capsule Endoscopy Dataset. *Electronics* **2022**, *11*, doi:10.3390/electronics11172747.
14. Hosain, A.S.; Islam, M.; Mehedi, M.H.K.; Kabir, I.E.; Khan, Z.T. Gastrointestinal disorder detection with a transformer based approach. In Proceedings of the 2022 IEEE 13th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2022; pp. 0280-0285.
15. Lima, D.L.S.; Pessoa, A.C.P.; De Paiva, A.C.; da Silva Cunha, A.M.T.; Júnior, G.B.; De Almeida, J.D.S. Classification of Video Capsule Endoscopy Images Using Visual Transformers. In Proceedings of the 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), 2022; pp. 1-4.
16. Graves, A. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850* **2013**.
17. Sak, H.; Senior, A.W.; Beaufays, F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. **2014**.
18. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. Slowfast networks for video recognition. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2019; pp. 6202-6211.
19. Wang, X.; Xiong, X.; Neumann, M.; Piergiovanni, A.; Ryoo, M.S.; Angelova, A.; Kitani, K.M.; Hua, W. Attentionnas: Spatiotemporal attention cell search for video classification. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16, 2020; pp. 449-465.
20. Bertasius, G.; Torresani, L. Classifying, segmenting, and tracking object instances in video with mask propagation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020; pp. 9739-9748.
21. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2015; pp. 3156-3164.
22. Bertasius, G.; Wang, H.; Torresani, L. Is space-time attention all you need for video understanding. *arXiv preprint arXiv:2102.05095* **2021**, *2*, 4.
23. Aoki, T.; Yamada, A.; Kato, Y.; Saito, H.; Tsuboi, A.; Nakada, A.; Niikura, R.; Fujishiro, M.; Oka, S.; Ishihara, S.; et al. Automatic detection of blood content in capsule endoscopy images based on a deep convolutional neural network. *J Gastroenterol Hepatol* **2020**, *35*, 1196-1200, doi:10.1111/jgh.14941.
24. Klang, E.; Barash, Y.; Margalit, R.Y.; Soffer, S.; Shimon, O.; Albshesh, A.; Ben-Horin, S.; Amitai, M.M.; Eliakim, R.; Kopylov, U. Deep learning algorithms for automated detection of Crohn's disease ulcers by video capsule endoscopy. *Gastrointest Endosc* **2020**, *91*, 606-613 e602, doi:10.1016/j.gie.2019.11.012.
25. Contributors, M. Openmmlab's next generation video understanding toolbox and benchmark. **2020**.
26. Alaskar, H.; Hussain, A.; Al-Aseem, N.; Liatsis, P.; Al-Jumeily, D. Application of Convolutional Neural Networks for Automated Ulcer Detection in Wireless Capsule Endoscopy Images. *Sensors (Basel)* **2019**, *19*, doi:10.3390/s19061265.
27. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017; pp. 1251-1258.

28. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934* **2020**.
29. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Gool, L.V. Temporal segment networks: Towards good practices for deep action recognition. In Proceedings of the European conference on computer vision, 2016; pp. 20-36.
30. Leenhardt, R.; Li, C.; Le Mouel, J.P.; Rahmi, G.; Saurin, J.C.; Cholet, F.; Boureille, A.; Amiot, X.; Delvaux, M.; Duburque, C.; et al. CAD-CAP: a 25,000-image database serving the development of artificial intelligence for capsule endoscopy. *Endosc Int Open* **2020**, *8*, E415-E420, doi:10.1055/a-1035-9088.
31. Aoki, T.; Yamada, A.; Aoyama, K.; Saito, H.; Tsuboi, A.; Nakada, A.; Niikura, R.; Fujishiro, M.; Oka, S.; Ishihara, S.; et al. Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network. *Gastrointest Endosc* **2019**, *89*, 357-363 e352, doi:10.1016/j.gie.2018.10.027.
32. Ding, Z.; Shi, H.; Zhang, H.; Meng, L.; Fan, M.; Han, C.; Zhang, K.; Ming, F.; Xie, X.; Liu, H.; et al. Gastroenterologist-Level Identification of Small-Bowel Diseases and Normal Variants by Capsule Endoscopy Using a Deep-Learning Model. *Gastroenterology* **2019**, *157*, 1044-1054 e1045, doi:10.1053/j.gastro.2019.06.025.