electronics

Article

Transformer-Based Disease Identification for Small-Scale Imbalanced Capsule Endoscopy Dataset

Long Bai [1], Liangyu Wang [1], Tong Chen [2,3], Yuanhao Zhao [1] and Hongliang Ren [1,4,5,6,*]

[1] Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong 999077, China
[2] School of Instrument Science and Opto-Electronics Engineering, Beijing Information Science and Technology University, Beijing 100101, China
[3] School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, China
[4] Shun Hing Institute of Advanced Engineering, The Chinese University of Hong Kong, Hong Kong 999077, China
[5] Department of Biomedical Engineering, National University of Singapore, Singapore 117583, Singapore
[6] NUS (Suzhou) Research Institute, Suzhou 215000, China
* Correspondence: hlren@ieee.org

Abstract: Vision Transformer (ViT) is emerging as a new leader in computer vision with its outstanding performance in many tasks (e.g., ImageNet-22k, JFT-300M). However, the success of ViT relies on pretraining on large datasets. It is difficult for us to use ViT to train from scratch on a small-scale imbalanced capsule endoscopic image dataset. This paper adopts a Transformer neural network with a spatial pooling configuration. Transfomer's self-attention mechanism enables it to capture long-range information effectively, and the exploration of ViT spatial structure by pooling can further improve the performance of ViT on our small-scale capsule endoscopy dataset. We trained from scratch on two publicly available datasets for capsule endoscopy disease classification, obtained 79.15% accuracy on the multi-classification task of the Kvasir-Capsule dataset, and 98.63% accuracy on the binary classification task of the Red Lesion Endoscopy dataset.

Keywords: vision transformer; wireless capsule endoscopy; disease identification; small-scale and imbalance dataset; pooling

Citation: Bai, L.; Wang, L.; Chen, T.; Zhao, Y.; Ren, H. Transformer-Based Disease Identification for Small-Scale Imbalanced Capsule Endoscopy Dataset. *Electronics* **2022**, *11*, 2747. https://doi.org/10.3390/electronics11172747

Academic Editor: Kah Phooi Seng

Received: 7 July 2022
Accepted: 27 August 2022
Published: 31 August 2022

1. Introduction

Gastrointestinal (GI) cancer is the second most deadly cancer worldwide [1], accounting for 35% of cancer-related deaths [2]. Endoscopy and radiological techniques can provide physicians with visual signals of the GI environment to diagnose. Compared to endoscopic techniques, radiological techniques produce inferior images with less discernible features. However, conventional endoscopes can only be inserted into the stomach and large intestine, but the small bowel structure is often overlapping and has no fixed shape, making it difficult for conventional endoscopes to penetrate. It was challenging to examine the small bowel until wireless capsule endoscopy (WCE) was developed [3,4]. Capsule endoscopy can easily access the small intestine and provide direct, clear video signals for physicians to observe possible lesions in the small bowel, such as ulcers, erosions, angiodysplasias, lymphoma, etc. [5,6].

WCE is a particular endoscope with imaging equipment, batteries, light source, and a signal-transmitting component. It is so tiny that a patient can directly swallow it. The WCE can move passively within a patient's gastrointestinal environment or selectively move and examine the patient's lesions under external control (e.g., magnetic control). Then, WCE will output a large amount of video information for the physicians' examination to determine the disease and follow-up treatment plan. However, the frame-by-frame analysis of the WCE output video is quite tedious. The examiner needs to be knowledgeable and experienced in small bowel disease and be focused at all times to identify the many different lesions [7,8]. These difficulties lead to high missing and misdiagnosis rates when the physicians manually deal with the WCE video information [9,10]. The WCE system overview can be found in Figure 1. A particular signal receiver will receive the image information transmitted by the capsule endoscopy inside the patient. Subsequently, a system based on artificial intelligence (AI) algorithms will analyze the acquired images in real time, thereby assisting physicians in making a diagnosis as quickly as possible.

With the rapid development of automation and visual analysis technology, deep learning (DL) has evolved and significantly grown in the last decade. DL methodologies have been widely used in various fields [11–18]. More and more DL algorithms are applied in automatic WCE diagnosis and analysis, and they have achieved excellent results. DL algorithms can bring very high diagnostic accuracy while helping to reduce the labor of physicians [19,20]. Meanwhile, some researchers apply DL to depth and motion estimation of WCE [21,22]. DL methodologies have shown great potential in WCE data, and will provide excellent help and improvement to future WCE technology.

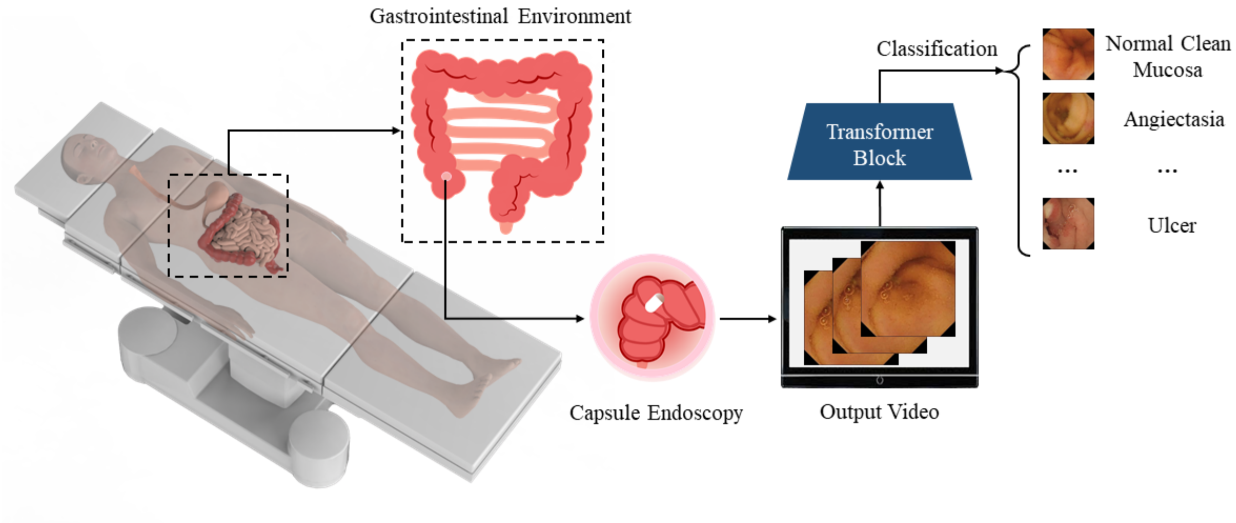


Figure 1. Overview of the capsule endoscopy system. The capsule endoscope collects images in the GI environment and outputs them in real time, and then AI algorithms process and classify the images.

The transformer is a recent emerging DL architecture [23]. Its architecture consists entirely of the self-attention mechanism and achieves better performance than Convolutional Neural Network (CNN). A. Dosovitskiy et al. [24] introduced the transformer into the task of image classification. Vision Transformer (ViT) showed extraordinary potential in image processing. However, the training of transformers often requires a large amount of data, and its excellent performance comes from pretraining based on large datasets. Even ImageNet [25] sometimes fails to unleash the full performance of ViT [24]. This reliance on pretraining is explained as locality inductive bias [26], requiring more data to learn features and visual representations. However, the datasets collected by capsule endoscopy usually have the following three major problems:

1. The amount of data is small;
2. The images of the dataset lack distinguishable features;
3. The data representing different diseases are not balanced. Common diseases may be represented in a large number of images, while for some rare diseases, there are only a few images for training.

This paper adopts a transformer network that imitates the existing pooling architecture. It can be trained from scratch and applied to the small and imbalanced capsule endoscopy dataset. We conduct experiments on the Kvasir-Capsule dataset [27] and Red Lesion Endoscopy (RLE) dataset [28], obtaining results that outperform existing baselines.

The rest of this paper is organized as follows: Sections 2 and 3 will introduce the related work and adopted methodology, respectively. Section 4 shows our experiments and results, followed by our discussion in Section 5. Finally, we conclude in Section 6.

2. Related Work

2.1. WCE Pathological Diagnosis

After the excogitating of WCE, pathological diagnosis methods regarding WCE have been seeking surmounts. In the early stage, researchers worked on developing publicly available WCE datasets. Several datasets, such as Kvasir-Capsule [27], RLE [28], KID [29], GIANA [30], have been published for researchers to reference and verification. Machine Learning methods have been widely developed for disease classification and detection in WCE images, such as support vector machines (SVM) [31] and single-shot multiBox detector (SSD) [32]. H. L. Gjestang et al. [33] presented a semi-supervised teacher–student framework with better performance than traditional supervised learning-based models—they utilized unlabeled samples for self-adjustment. As a result of the current popularity of self-attention mechanisms in computer vision, P. Muruganantham et al. [34] proposed a dual branch CNN model using self-attention mechanisms to bring out accuracy for WCE to classification and lesion localization. Traditional supervised DL methods require extensive samples to train and verify. However, obtaining large and specific lesion category samples from realistic clinic practice is challenging. R. Khadka et al. [35] proposed a new implicit model-agnostic meta-learning (iMAML) algorithm, which only requires small training samples, but performs high accuracy on unseen datasets.

Nevertheless, based on our discussion in Section 1, the most critical problems for the WCE datasets are the lack of distinguishable features, imbalanced classes, and the small amount of data. The commonly used convolutional calculation has space constraints, and the network's attention area is relatively limited. ViT can learn the most appropriate inductive bias according to the task objective and the layer's position in the network. Furthermore, the multi-head mechanism ensures that ViT can focus on multiple discriminative parts. These advantages allow ViT to achieve better results on WCE tasks.

2.2. Vit Training from Scratch

Several approaches have been proposed to overcome the problem that ViT must rely on large datasets for pretraining. Data-efficient Image Transformer (DeiT) [36] introduced a ViT-based teacher–student framework that allows student models to learn from teacher models through an attention mechanism. This ViT-based knowledge distillation significantly reduced training time and computational resource consumption. Tokens-To-Token (T2T) [37] learned the deep-narrow structure of CNNs and developed a layerwise Tokens-to-Token transformation methodology to learn the local structure around tokens. Convolutional vision Transformer (CvT) [38] attempted to bring the CNN's ability to extract low-level information to ViT. It used Convolutional Token Embedding and mimicked CNN's process of learning features, which combined the benefits of both CNN and ViT. Swin Transformer [39] merged image patches to construct a hierarchical Transformer, and used attention computed in a non-overlapping local window. Class-Attention in Image Transformer (CaiT) [40] proposed LayerScale to make the convergence of Deep ViT easier and used class-attention layers to make the processing of class tokens more efficient. Y. Liu et al. [41] utilized a self-supervised architecture for learning spatial information in a single image. It can work in conjunction with supervised learning of ViT, allowing ViT to be more robust and accurate when trained on small-size datasets. S. H. Lee et al. [42] proposed two generic add-on modules, Shifted Patch Tokenization and Locality Self-Attention, to solve the problem of the locality inductive bias.

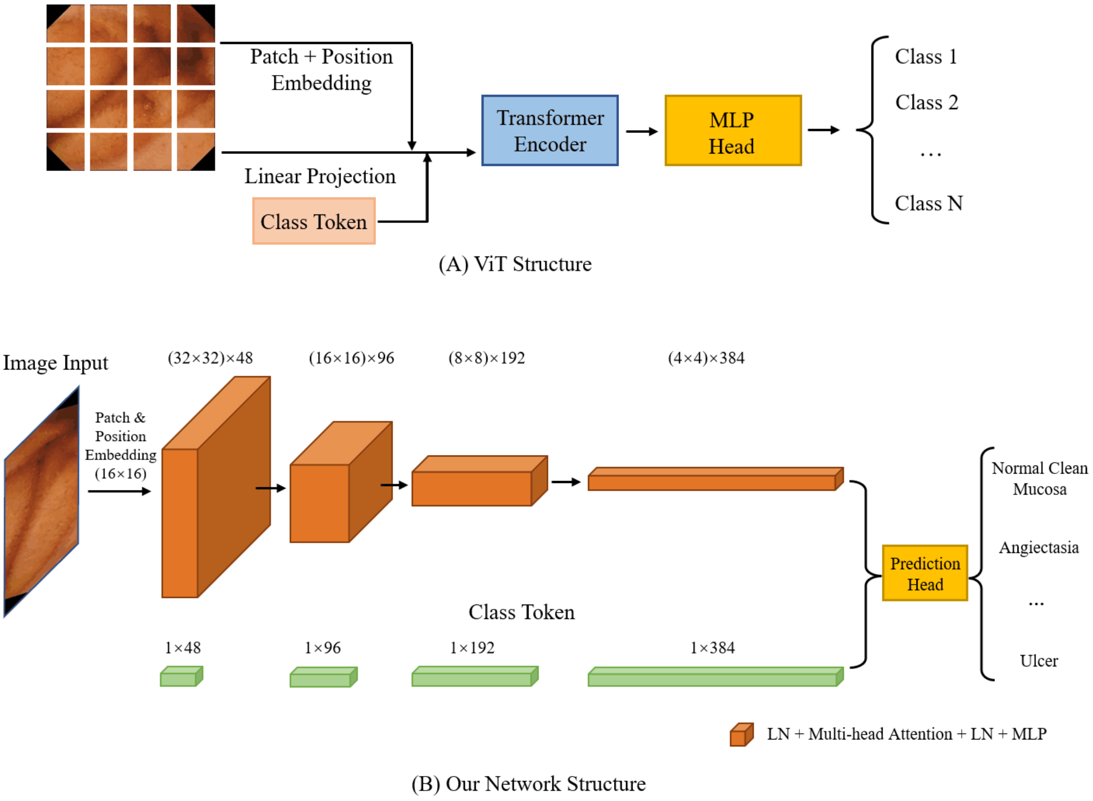
However, despite the tremendous progress in training from scratch on natural datasets, ViT, as a state-of-the-art technique, has not been popularized in processing WCE data.

Compared with the tiny dataset commonly used in general computer vision, the WCE dataset has fewer data and is highly imbalanced. Thus, it is not easy to directly port existing methods for handling small-size datasets, hindering the further promotion and development of ViT on the WCE dataset.

3. Methodology

3.1. Vision Transformer

ViT combines knowledge from computer vision and natural language processing. The original ViT [24] consists of patch embedding, position embedding, transformer encoder, class token, and classification head. The patch embedding divides the original image into 16 × 16 patches and flattens the patches into a 1 × n sequence. Position embedding provides the model with the location information of the patch so the model does not need to learn patch stitching based on the semantics of the patches, saving the learning cost. The transformer encoder is used to process the flattened patches. It has LayerNorm (LN), multi-head attention, LN, multi-layer perceptron (MLP) in order. The transformer encoder is established based on the multi-head self-attention mechanism, which assigns different weights to the input information to aggregate information. ViT builds the network by stacking multiple transformer encoder blocks. After being processed by the transformer encoder, the patches are still 1 × n matrices. The class token is a separate vector used to aggregate the image feature of the corresponding processed patch. When connecting the class token to the linear classifier, the classification-predicted results can be obtained. Figure 2A shows the typical architecture of ViT.



(A) ViT Structure

(B) Our Network Structure

Figure 2. The neural network architecture. (**A**) An original ViT structure for the classification task. The image is first converted into flattened patches through Patch Embedding and Position Embedding, then processed by the Transformer Encoder. The prediction result is obtained after the MLP Head. (**B**) The spatial structure of the original ViT keeps unchanged, while we adopt a ViT structure with 3-step pooling. $(n \times n) \times m$ represents a $(n^2 \times m)$ 2D matrix when in ViT and a $(n \times n \times m)$ 3D matrix when conducting pooling operation.

3.2. Pooling in CNN

Pooling is a commonly used module in CNNs, which is also known as downsampling [43–45]. It is usually used after the convolutional layer to reduce the feature map dimension. It can effectively reduce the network parameters, prevent overfitting, improve model robustness, and reduce information redundancy.

Max Pooling and Average Pooling are the two most commonly used pooling methods. Max Pooling selects the maximum pixel value in the image as the value after Pooling. The gradient is backpropagated through the maximum value of the forward propagation process, and the gradient at other locations is 0. Max Pooling is more concerned with the local features in the image. Average Pooling involves selecting the average value in the image region as the value after pooling, which focuses more on the global features of the image. Mix Pooling is a combination of Max Pooling and Average Pooling, and it also has the complementary advantage of the two methods. Researchers utilize Concat operation and Add operation to combine Max Pooling and Average Pooling. ResNet [43] uses the Add pooling operation, which adds the corresponding feature maps and then performs convolution operations. The corresponding feature maps share a convolution kernel. DenseNet [44] adopts Concat pooling operation. Each channel corresponds to one convolutional kernel with higher computational consumption. All these network architectures show a spatially reduced dimensionality design.

3.3. Pooling in ViT

In the original ViT, the spatial dimension of each layer remains unchanged, and each layer has the same number of tokens. Inspired by [46], the pooling operation in ViT can effectively improve the model's generalization ability. We can learn from the network architecture in CNN to obtain a downsampling ViT architecture in the spatial dimension.

However, the original ViT-processed data are presented as 2D matrix data, so we cannot directly apply pooling operation to them. B. Heo et al. [46] present pooling layers in ViT. They first reshape the 2D matrix back to 3D at the beginning of the pooling layer. Thus, the depthwise convolution can be performed here to conduct the common pooling operation in CNN, with the variation of the space dimension and number of channels. Then, the 3D data will be reshaped again into the 2D matrix, which can be added back into the ViT module. Moreover, the class token is spatially aligned with ViT's pooling module using fully connected layer extensions. Figure 2B shows the detailed spatial size of our network architecture. We follow the same setup in the original ViT and only modify the spatial dimensions in the transformer encoder. We design the whole structure in the style of 3-step pooling so that the spatial dimension will be reduced three times. The final prediction result will be obtained through the classifier head after three consecutive downsamplings of the processed data.

4. Experiment Validation

4.1. Datasets

The experiments were conducted based on the Kvasir-Capsule dataset [27] and RLE dataset [28]. Their image examples and data distribution are attached in Figures A1 and A2 and Tables A1 and A2. In the Kvasir-Capsule dataset, we remove the three classes from the Anatomy category, and retain the Luminal findings category. Thus, it has 11 classes, including 10 different categories of diseases (Angiectasia, Blood-fresh, Blood-hematin, Erosion, Erythema, Foreign Body, Pylorus, Lymphangiectasia, Ulcer, Reduced Mucosal View) and Normal Clean Mucosa images. The image resolution is 336 × 336. The data distribution of the Kvasir-Capsule dataset is very imbalanced—the largest class has 34,338 images, and the smallest class has only 12 images. The RLE dataset was originally for tissue hemorrhage segmentation in capsule endoscopy images. Its image resolution is 320 × 320. We rearranged the dataset into two classes, normal and bleeding images, and performed a binary classification task on this dataset.

Meanwhile, to ensure the robustness of prediction and prevent over-fitting, data augmentation strategies such as random rotation, flipping, and random noise are used to expand the datasets and improve their diversity and generalization ability. The training and testing sets are divided by the ratio of 4:1.

4.2. Model Evaluation

The neural networks were completed based on the Python PyTorch framework. The training and validation were conducted on a server with Ubuntu 18.04 LTS operating system, 2 NVIDIA GeForce RTX 3090 GPUs, and an AMD Ryzen™ 9 5950X CPU. We set the batch size to 128, and all models were trained for 100 epochs. The learning rate is set to 0.001 with the Adam optimizer. The comparison algorithms can be split into CNNs and ViTs as follows:

- CNNs: ResNet32 [43], ResNet56 [43], VGGNet19 [47], WRN16-2 [48], WRN40-2 [48], MobileNetV2 [49], ShuffleNetV1 [50], ShuffleNetV2 [51];
- ViTs: Swin-Transformer [39], CaiT [40].

We implemented all the experiments in our environment. Table 1 and Figure 3 show the classification accuracy and comparison. Firstly, among all CNN algorithms, ShuffleNetV1 achieves very excellent results on both datasets. ResNet56 achieves the highest accuracy among all CNNs on the RLE dataset, a binary classification task, but slightly lower than ShuffleNetV1 on the Kvasir-Capsule dataset, a multiclassification task. Secondly, in ViT algorithms, both CaiT and Swin Transformer show superior performance to CNN on both datasets. However, CaiT features a performance that exceeds Swin Transformer on the Kvasir-Capsule dataset, while Swin Transformer defeats CaiT on the RLE dataset. Lastly, the method we employed outperforms both CaiT and Swin Transformer quite clearly on both WCE datasets, with the accuracy of 79.15% on the Kvasir-Capsule dataset and 98.63% on the RLE dataset. We also present the inference time of different methodologies for a single image. ViT-based methods are slightly slower, but in real-time diagnostic applications, a difference of 0.06 ms (i.e., 6×10^{-5} s) does not have a substantial impact.

Table 1. The results of comparison experiments.

Models	Kvasir-Capsule	RLE	Inference Time/Per Image
ResNet32 [43]	74.49%	95.22%	0.41 ms
ResNet56 [43]	74.76%	95.77%	0.42 ms
VGGNet19 [47]	74.77%	94.30%	0.41 ms
WRN16_2 [48]	74.52%	93.01%	0.41 ms
WRN40_2 [48]	74.50%	93.38%	0.42 ms
MobileNetV2 [49]	74.91%	94.67%	0.41 ms
ShuffleNetV1 [50]	75.62%	95.40%	0.42 ms
ShuffleNetV2 [51]	75.33%	94.49%	0.41 ms
Swin [39]	77.88%	98.33%	0.47 ms
CaiT [40]	78.63%	97.26%	0.48 ms
Ours	79.15%	98.63%	0.47 ms

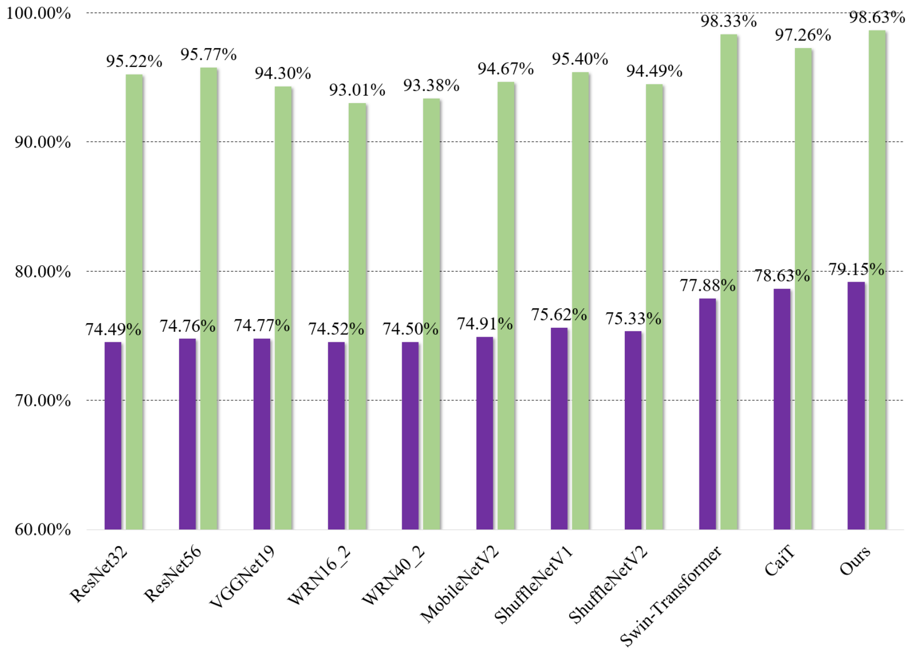


Figure 3. The prediction accuracy comparison of all algorithms. The purple blocks show the prediction accuracy on the multi-classification task of the Kvasir-Capsule dataset, and the light green blocks show the prediction accuracy on the binary classification task of the RLE dataset. Our method exhibits the best performance on both tasks.

5. Discussion

This section compares our methodology with other related and recently published WCE classification solutions. The associated solutions for WCE classification are listed in Table 2.

Table 2. Comparison of the classification results of this study and other studies.

Authors	Methodology	Dataset	Accuracy
M. Sharif et al. [52]	Geometric Features, CNN	10 WCE videos	99.1%
F. Rustam et al. [53]	MobileNet, BIR	Bleeding Detection, Binary Classification	99.3%
X. Zhao et al. [54]	CNN Backbone, LSTM, Transformer	113 WCE videos	93.0%
H. L. Gjestang et al. [33]	Teacher–student Framework	Kvasir-Capsule [27]	69.5%
H. L. Gjestang et al. [33]	Teacher–student Framework	HyperKvasir [55]	89.3%
P. Muruganantham et al. [34]	Self-attention CNN	Processed Kvasir-Capsule [27]	95.4%
P. Muruganantham et al. [34]	Self-attention CNN	Processed RLE [28]	95.1%
S. Biradher et al. [56]	CNN	Processed RLE	98.5%
D. Bajhaiya et al. [57]	DenseNet121	Ulcer Classification [27]	99.9%
N. Goel et al. [58]	Dilated Input Context Retention CNN	8 WCE videos	96.0%
N. Goel et al. [58]	Dilated Input Context Retention CNN	KID [29]	93.0%
A. Srivastava et al. [59]	FocalConvNet	Processed Kvasir-Capsule [27]	63.7%

Most of the research on WCE classification is based on CNN and its variants. In addition to exploring the existing CNN architectures [53,56,57], F. Rustam et al. [53] and N. Goel et al. [58] tried to improve the spatial architecture of the original CNN. A. Srivastavaet et al. [59] used the Focal Modulation Guided method to integrate global context information to CNN. M. Sharif et al. [52], P. Muruganantham et al. [34], X. Zhao et al. [54] added external information or components to CNN to improve the prediction ability. H. L. Gjestang et al. [33] explored the semi-supervised teacher–student framework to overcome the difficulties and deficiencies of annotation in medical data. Most of the above methods are based on CNN attempts, while X. Zhao et al. [54] introduced a transformer to process the WCE videos. However, they still used CNN as the feature extractor while utilizing the transformer to process temporal features and capture the interframe dependencies in videos.

Our methodology is based on a purely ViT network without the convolutional operations from the CNN. We employ spatial downsampling to integrate the spatial structure of ViT.

M. Raghu et al. [60] dissects the difference between CNN and ViT in terms of receptive field in detail. Usually, CNNs have only a fixed-size kernel (size 3 or 7), so CNNs cannot obtain global perception in the initial layer. Instead, the CNN obtains the local receptive field at initial layers and then gradually expands the receptive field by repeatedly "convolving" the information around the kernel layer by layer. Therefore, researchers must keep stacking convolutional layers to obtain a sizable receptive field via CNN. In contrast, based on the self-attentive mechanism of ViT, the receptive field (i.e., the mean attention distance of the self-attention head) of ViT in the initial layers has the local receptive field like CNN, but also has the global field of view. ViT's global and long-range perception will ensure that when ViT makes the decision, the features ViT has used already cover sufficient information from the original image. Obviously, this also brings better performance to our disease classification for WCE. From Table 1, we know that the CaiT [40] and Swin Transformer [39] already outperform all CNN backbones. We also adopted spatial pooling to integrate the spatial structure of ViT. We can think of this as ViT learning from CNN to build the network structure, and then explore the potential of ViT on the WCE dataset. Excitingly, our method successfully achieves the best performance, and this proves the usability of ViT on the WCE dataset. Further work on classification and diagnosis on the WCE dataset will aim to reduce the number of parameters of ViT and speed up the inference speed, so that ViT can be better implemented in the real-world scenario of clinical medical treatment.

6. Conclusions

This paper presents a ViT-based WCE image disease classification algorithm that can be trained from scratch. We adopt several CNN and ViT algorithms for comparison experiments. The experimental results show that, in terms of accuracy, our adopted methodology surpasses all comparison methods in processing small-scale imbalanced WCE datasets, demonstrating the feasibility of applying the ViT algorithm directly to small-scale medical datasets. Our proposed method can derive classification results with the inference time of 1 ms, making it suitable for real-time diagnosis. However, although our model achieves greater accuracy than existing methods on two public datasets, the slightest misdiagnosis or miss in medical applications can be fatal. Thus, our solution can only be used for prescreening before the doctor's manual diagnosis, or for review after the doctor's diagnosis. AI still needs further exploration in actual medical diagnosis and treatment.

Author Contributions: Conceptualization, L.B. and H.R.; methodology, L.B. and L.W.; software, L.B. and L.W.; validation, L.B. and L.W.; formal analysis, L.B.; investigation, L.B.; resources, H.R.; data curation, L.B., T.C. and Y.Z.; writing—original draft preparation, L.B.; writing—review and editing, L.B., L.W. and H.R.; visualization, L.B.; supervision, H.R.; project administration, H.R.; funding acquisition, H.R. All authors have read and agreed to the published version of the manuscript.

Funding: National Key R&D Program of China from the Ministry of Science and Technology (MOST) of China: Grant 2018YFB1307700 (also with subprogram 2018YFB1307703); Shun Hing Institute of Advanced Engineering at The Chinese University of Hong Kong (CUHK): BME-p1-21/8115064; Hong Kong Research Grants Council (RGC): Collaborative Research Fund (CRF C4026-21GF and CRF C4063-18G) and General Research Fund (GRF #14211420); Shenzhen-Hong Kong-Macau Technology Research Programme (Type C): Grant 202108233000303.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

BIR	Bleedy Image Recognition
CaiT	Class-Attention in Image Transformer
CvT	Convolutional vision Transformer
DeiT	Data-efficient Image Transformer
DL	Deep Learning
GI	Gastrointestinal
GIANA	Gastrointestinal Image ANAlysis
LN	LayerNorm
MLP	Multilayer Perceptron
RLE	Red Lesion Endoscopy
SSD	Single-shot MultiBox Detector
SVM	Support Vector Machine
T2T	Tokens-To-Token
ViT	Vision Transformer
WCE	Wireless Capsule Endoscopy

Appendix A

This section presents the image examples and data distribution in the Kvasir-Capsule Dataset [27] and RLE Dataset [28].

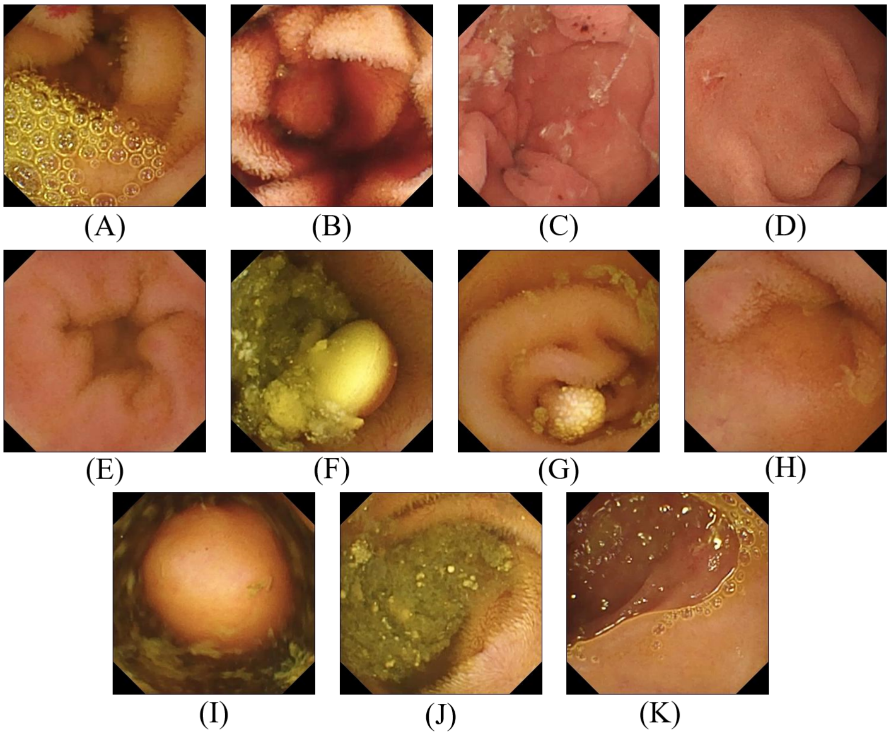


Figure A1. Examples of images in the Kvasir-Capsule dataset [27]. (**A**) Angiectasia; (**B**) Blood-fresh; (**C**) Blood-hematin (**D**) Erosion; (**E**) Erythema; (**F**) Foreign Body; (**G**) Lymphangiectasia; (**H**) Normal Clean Mucosa; (**I**) Pylorus; (**J**) Reduced Mucosal View; (**K**) Ulcer.

Table A1. Data distribution of Kvasir-Capsule dataset [27]. The eleven categories of this dataset show strong imbalances.

Category	Number of Images
Normal Clean Mucosa	34,338
Reduced Mucosal View	2906
Lymphangiectasia	592
Erythema	159
Angiectasia	866
Blood—fresh	446
Blood—Hematin	12
Erosion	506
Ulcer	854
Polyp	55
Foreign Body	776

Table A2. Data distribution of RLE dataset [28].

Category	Number of Images
Normal	2160
Bleeding	1125

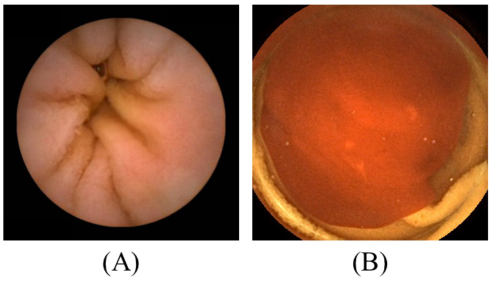


Figure A2. Examples of images in the RLE dataset [28]. (**A**) Normal; (**B**) Bleeding.

References

1. Arnold, M.; Abnet, C.C.; Neale, R.E.; Vignat, J.; Giovannucci, E.L.; McGlynn, K.A.; Bray, F. Global burden of 5 major types of gastrointestinal cancer. *Gastroenterology* **2020**, *159*, 335–349. [CrossRef] [PubMed]
2. Center, M.; Siegel, R.; Jemal, A. *Global Cancer Facts & Figures*; American Cancer Society: Atlanta, GA, USA, 2011; Volume 3, p. 52.
3. Flemming, J.; Cameron, S. Small bowel capsule endoscopy: Indications, results, and clinical benefit in a University environment. *Medicine* **2018**, *97*, e0148. [CrossRef] [PubMed]
4. Aktas, H.; Mensink, P.B. Small bowel diagnostics: Current place of small bowel endoscopy. *Best Pract. Res. Clin. Gastroenterol.* **2012**, *26*, 209–220. [CrossRef] [PubMed]
5. McLaughlin, P.D.; Maher, M.M. Primary malignant diseases of the small intestine. *Am. J. Roentgenol.* **2013**, *201*, W9–W14. [CrossRef] [PubMed]
6. Thomson, A.; Keelan, M.; Thiesen, A.; Clandinin, M.; Ropeleski, M.; Wild, G. Small bowel review: Diseases of the small intestine. *Dig. Dis. Sci.* **2001**, *46*, 2555–2566. [CrossRef]
7. Zheng, Y.; Hawkins, L.; Wolff, J.; Goloubeva, O.; Goldberg, E. Detection of lesions during capsule endoscopy: Physician performance is disappointing. *Off. J. Am. Coll. Gastroenterol. ACG* **2012**, *107*, 554–560. [CrossRef]
8. Chetcuti Zammit, S.; Sidhu, R. Capsule endoscopy–recent developments and future directions. *Expert Rev. Gastroenterol. Hepatol.* **2021**, *15*, 127–137. [CrossRef]
9. Rondonotti, E.; Soncini, M.; Girelli, C.M.; Russo, A.; Ballardini, G.; Bianchi, G.; Cantù, P.; Centenara, L.; Cesari, P.; Cortelezzi, C.C.; et al. Can we improve the detection rate and interobserver agreement in capsule endoscopy? *Dig. Liver Dis.* **2012**, *44*, 1006–1011. [CrossRef]
10. Kaminski, M.F.; Regula, J.; Kraszewska, E.; Polkowski, M.; Wojciechowska, U.; Didkowska, J.; Zwierko, M.; Rupinski, M.; Nowacki, M.P.; Butruk, E. Quality indicators for colonoscopy and the risk of interval cancer. *N. Engl. J. Med.* **2010**, *362*, 1795–1803. [CrossRef]
11. Shen, D.; Wu, G.; Suk, H.I. Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* **2017**, *19*, 221. [CrossRef]
12. Wang, A.; Islam, M.; Xu, M.; Ren, H. Rethinking Surgical Instrument Segmentation: A Background Image Can Be All You Need. *arXiv* **2022**, arXiv:2206.11804.

13. Bai, L.; Chen, S.; Gao, M.; Abdelrahman, L.; Al Ghamdi, M.; Abdel-Mottaleb, M. The Influence of Age and Gender Information on the Diagnosis of Diabetic Retinopathy: Based on Neural Networks. In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Mexico, 1–5 November 2021; pp. 3514–3517. Available online: https://embc.embs.org/2021/ (accessed on 30 November 2021).
14. Bai, L.; Yang, J.; Wang, J.; Lu, M. An Overspeed Capture System Based on Radar Speed Measurement and Vehicle Recognition. In Proceedings of the International Conference on International Conference on Artificial Intelligence for Communications and Networks, Virtual Event, 19–20 December 2020; pp. 447–456.
15. Kim, H.; Park, J.; Lee, H.; Im, G.; Lee, J.; Lee, K.B.; Lee, H.J. Classification for Breast Ultrasound Using Convolutional Neural Network with Multiple Time-Domain Feature Maps. *Appl. Sci.* **2021**, *11*, 10216. [CrossRef]
16. Jang, Y.; Jeong, I.; Cho, Y.K. Identifying impact of variables in deep learning models on bankruptcy prediction of construction contractors. In *Engineering, Construction and Architectural Management*; Emerald Publishing Limited: Bradford, UK, 2021.
17. Kang, S.H.; Han, J.H. Video captioning based on both egocentric and exocentric views of robot vision for human-robot interaction. *Int. J. Soc. Robot.* **2021**, 1–11. [CrossRef]
18. Che, H.; Jin, H.; Chen, H. Learning Robust Representation for Joint Grading of Ophthalmic Diseases via Adaptive Curriculum and Feature Disentanglement. *arXiv* **2022**, arXiv:2207.04183.
19. Yuan, Y.; Meng, M.Q.H. Deep learning for polyp recognition in wireless capsule endoscopy images. *Med. Phys.* **2017**, *44*, 1379–1389. [CrossRef]
20. Karargyris, A.; Bourbakis, N. Detection of small bowel polyps and ulcers in wireless capsule endoscopy videos. *IEEE Trans. Biomed. Eng.* **2011**, *58*, 2777–2786. [CrossRef]
21. Li, L.; Li, X.; Yang, S.; Ding, S.; Jolfaei, A.; Zheng, X. Unsupervised-learning-based continuous depth and motion estimation with monocular endoscopy for virtual reality minimally invasive surgery. *IEEE Trans. Ind. Inform.* **2020**, *17*, 3920–3928. [CrossRef]
22. Ozyoruk, K.B.; Gokceler, G.I.; Bobrow, T.L.; Coskun, G.; Incetan, K.; Almalioglu, Y.; Mahmood, F.; Curto, E.; Perdigoto, L.; Oliveira, M.; et al. EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. *Med. Image Anal.* **2021**, *71*, 102058. [CrossRef]
23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
24. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
25. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
26. Neyshabur, B. Towards learning convolutions from scratch. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 8078–8088.
27. Smedsrud, P.H.; Thambawita, V.; Hicks, S.A.; Gjestang, H.; Nedrejord, O.O.; Næss, E.; Borgli, H.; Jha, D.; Berstad, T.J.D.; Eskeland, S.L.; et al. Kvasir-Capsule, a video capsule endoscopy dataset. *Sci. Data* **2021**, *8*, 142. [CrossRef] [PubMed]
28. Coelho, P.; Pereira, A.; Salgado, M.; Cunha, A. A deep learning approach for red lesions detection in video capsule endoscopies. In Proceedings of the International Conference Image Analysis and Recognition, Póvoa de Varzim, Portugal, 27–29 June 2018; pp. 553–561.
29. Koulaouzidis, A.; Iakovidis, D.K.; Yung, D.E.; Rondonotti, E.; Kopylov, U.; Plevris, J.N.; Toth, E.; Eliakim, A.; Johansson, G.W.; Marlicz, W.; et al. KID Project: An internet-based digital video atlas of capsule endoscopy for research purposes. *Endosc. Int. Open* **2017**, *5*, E477–E483. [CrossRef] [PubMed]
30. Bernal, J.; Aymeric, H. Gastrointestinal Image Analysis (GIANA) Angiodysplasia d&l Challenge. Web-page of the 2017 Endoscopic Vision Challenge. 2017. Available online: https://endovissub2017-giana.grand-challenge.org/ (accessed on 20 May 2018).
31. Amiri, Z.; Hassanpour, H.; Beghdadi, A. A Computer-Aided Method for Digestive System Abnormality Detection in WCE Images. *J. Healthc. Eng.* **2021**, *2021*, 7863113. [CrossRef] [PubMed]
32. Saito, H.; Aoki, T.; Aoyama, K.; Kato, Y.; Tsuboi, A.; Yamada, A.; Fujishiro, M.; Oka, S.; Ishihara, S.; Matsuda, T.; et al. Automatic detection and classification of protruding lesions in wireless capsule endoscopy images based on a deep convolutional neural network. *Gastrointest. Endosc.* **2020**, *92*, 144–151. [CrossRef]
33. Gjestang, H.L.; Hicks, S.A.; Thambawita, V.; Halvorsen, P.; Riegler, M.A. A self-learning teacher-student framework for gastrointestinal image classification. In Proceedings of the 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), Aveiro, Portugal, 7–9 June 2021; pp. 539–544.
34. Muruganantham, P.; Balakrishnan, S.M. Attention aware deep learning model for wireless capsule endoscopy lesion classification and localization. *J. Med Biol. Eng.* **2022**, *42*, 157–168. [CrossRef]
35. Khadka, R.; Jha, D.; Hicks, S.; Thambawita, V.; Riegler, M.A.; Ali, S.; Halvorsen, P. Meta-learning with implicit gradients in a few-shot setting for medical image segmentation. *Comput. Biol. Med.* **2022**, *143*, 105227. [CrossRef] [PubMed]
36. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning, Virtual Event, 18–24 July 2021; pp. 10347–10357.
37. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.H.; Tay, F.E.; Feng, J.; Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 558–567.

38. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. Cvt: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 22–31.
39. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
40. Touvron, H.; Cord, M.; Sablayrolles, A.; Synnaeve, G.; Jégou, H. Going deeper with image transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 32–42.
41. Liu, Y.; Sangineto, E.; Bi, W.; Sebe, N.; Lepri, B.; Nadai, M. Efficient training of visual transformers with small datasets. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 23818–23830.
42. Lee, S.H.; Lee, S.; Song, B.C. Vision transformer for small-size datasets. *arXiv* **2021**, arXiv:2112.13492.
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
44. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
45. Wang, J.; Li, J.; Ding, L.; Wang, Y.; Xu, T. PAPooling: Graph-based Position Adaptive Aggregation of Local Geometry in Point Clouds. *arXiv* **2021**, arXiv:2111.14067.
46. Heo, B.; Yun, S.; Han, D.; Chun, S.; Choe, J.; Oh, S.J. Rethinking spatial dimensions of vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 11936–11945.
47. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
48. Zagoruyko, S.; Komodakis, N. Wide residual networks. *arXiv* **2016**, arXiv:1605.07146.
49. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
50. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6848–6856.
51. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 116–131.
52. Sharif, M.; Attique Khan, M.; Rashid, M.; Yasmin, M.; Afza, F.; Tanik, U.J. Deep CNN and geometric features-based gastrointestinal tract diseases detection and classification from wireless capsule endoscopy images. *J. Exp. Theor. Artif. Intell.* **2021**, *33*, 577–599. [CrossRef]
53. Rustam, F.; Siddique, M.A.; Siddiqui, H.U.R.; Ullah, S.; Mehmood, A.; Ashraf, I.; Choi, G.S. Wireless capsule endoscopy bleeding images classification using CNN based model. *IEEE Access* **2021**, *9*, 33675–33688. [CrossRef]
54. Zhao, X.; Fang, C.; Gao, F.; De-Jun, F.; Lin, X.; Li, G. Deep Transformers for Fast Small Intestine Grounding in Capsule Endoscope Video. In Proceedings of the 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), Nice, France, 13–16 April 2021; pp. 150–154.
55. Borgli, H.; Thambawita, V.; Smedsrud, P.H.; Hicks, S.; Jha, D.; Eskeland, S.L.; Randel, K.R.; Pogorelov, K.; Lux, M.; Nguyen, D.T.D.; et al. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Sci. Data* **2020**, *7*, 1–14. [CrossRef] [PubMed]
56. Biradher, S.; Aparna, P. Classification of Wireless Capsule Endoscopy Bleeding Images using Deep Neural Network. In Proceedings of the 2022 IEEE Delhi Section Conference (DELCON), Delhi, India, 11–13 February 2022; pp. 1–4.
57. Bajhaiya, D.; Unni, S.N. Deep learning-enabled classification of gastric ulcers from wireless-capsule endoscopic images. In *Medical Imaging 2022: Digital and Computational Pathology*; SPIE: San Diego, CA, USA, 2022; Volume 12039, pp. 352–356.
58. Goel, N.; Kaur, S.; Gunjan, D.; Mahapatra, S. Dilated CNN for abnormality detection in wireless capsule endoscopy images. *Soft Comput.* **2022**, *26*, 1231–1247. [CrossRef]
59. Srivastava, A.; Tomar, N.K.; Bagci, U.; Jha, D. Video Capsule Endoscopy Classification using Focal Modulation Guided Convolutional Neural Network. *arXiv* **2022**, arXiv:2206.08298.
60. Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; Dosovitskiy, A. Do vision transformers see like convolutional neural networks? *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12116–12128.