Endoscopic Image classification with Vision Transformers

Preeti Bissoonauth-Daiboo*
preetidaiboo@gmail.com
Department of Software and Information Systems, University of Mauritius
Reduit, Mauritius

Maleika Heenaye-Mamode Khan
m.mamodekhan@uom.ac.mu
Department of Software and Information Systems, University of Mauritius
Reduit, Mauritius

Muhammad Muzzammil Auzine
mmuzzammil.auzine@gmail.com
Department of Software and Information Systems, University of Mauritius
Reduit, Mauritius

Sunilduth Baichoo
sbaichoo@uom.ac.mu
Department of Software and Information Systems, University of Mauritius
Reduit, Mauritius

Xiaohong. Gao
x.gao@mdx.ac.uk
Department of Computer Science, Middlesex University
London, United Kingdom

Zaid Heetun
zaidheetun@gmail.com
Center for Gastroenterology and Hepatology, Dr Abdool Gaffoor Jeetoo Hospital
Port Louis, Mauritius

ABSTRACT

Convolutional Neural Networks (CNNs) have been the state-of-the-art techniques applied in the field of medical imaging for numerous image processing tasks. Recently, vision transformer networks are emerging as another technique, complementing current CNNs in the medical field providing on-par performance while also having a number of unique characteristics that may be useful for medical image processing. While CNNs have been predominantly applied to artefact detection and classification in endoscopic images, ViT has been sparsely applied in this area. Additionally, both CNN and ViT have been sparingly applied to colour misalignment artefact classification. In this work, we, therefore, explore the application of Vision Transformer (ViT) in the classification of artefacts in endoscopic images of the gastrointestinal tract organs. Furthermore, the performance of ViT is compared to that of CNN in the classification of colour misalignment artefacts. Our customised ViT model, based on DeiT (Data-efficient image Transformers), has obtained an accuracy of 96.33% as compared to the CNN based Inceptionv3 model with an accuracy of 78.67% and InceptionResNetv2 with 76.67%. The results demonstrate that when pretrained on ImageNet, ViT offer better performance than CNNs in colour misalignment artefact classification. This is due to the ability of ViT to better depict the relationship between image pixels through self-attention weights. Moreover, the built-in self-attention mechanism offers fresh insight into the decision-making processes of the model.

CCS CONCEPTS

• **Computing methodologies → Object recognition**.

KEYWORDS

Vision Transformer, CNN, Artefact Classification, Colour Misalignment

ACM Reference Format:
Preeti Bissoonauth-Daiboo, Maleika Heenaye-Mamode Khan, Muhammad Muzzammil Auzine, Sunilduth Baichoo, Xiaohong. Gao, and Zaid Heetun. 2023. Endoscopic Image classification with Vision Transformers. In *2023 The 7th International Conference on Advances in Artificial Intelligence (ICAAI) (ICAAI 2023), October 13–15, 2023, Istanbul, Turkiye.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3633598.3633623

1 INTRODUCTION

Esophageal cancer ranked seventh in terms of cancer incidence rate and sixth in overall cancer mortality rate worldwide in 2020 [17]. Endoscopy is a commonly performed imaging procedure for early screening of diseases in gastrointestinal tract organs. However, artefacts are caused by the rapid movement of the endoscope, the unique characteristics and constrained environment of the organs under examination. While there are various artefacts such as saturation, specularity, debris, bubbles and contrast [1], one prominent artefact in endoscopic images is colour misalignment. During endoscopy, the endoscopic camera takes pictures sequentially in red, blue and green, then these three channels are combined to form a colourful picture. Since the esophageal food pipe is of relatively small size, due to the fast movement of the camera, these three channel pictures may not be captured from the same spot causing non-realistic appearance, which is the colour misalignment artefact as illustrated in Figure 1 [6].

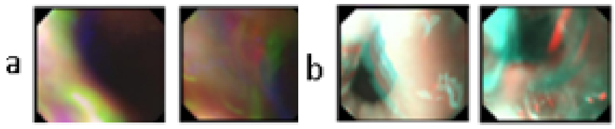

Figure 1: Colour misalignment (a - White Light Endoscopy, b- Narrow Band Imaging [6]

ICAAI 2023, October 13–15, 2023, Istanbul, Turkiye

ACM ISBN 979-8-4007-0898-5/23/10…$15.00
https://doi.org/10.1145/3633598.3633623

These artefacts hinder the detection of abnormal variations in the tract organs. These abnormal regions might eventually lead to development of cancers if undetected and untreated as they are mixed with objects of interests. In-depth knowledge, experience and training are required by endoscopists in order to identify the subtle changes and abnormalities in endoscopic images. There have been significant advances in imaging technology in the recent decades such as image-enhanced endoscopy and magnifying endoscopy to facilitate early detection of cancers [23] but the presence of artefacts in endoscopic images has often caused misdiagnosis. The rate of undetected upper gastrointestinal cancers over the past 3 years was high (25%) and the main reason was attributed to endoscopic errors [11].

With the enhancements in machine learning and deep learning, novel techniques using Convolution Neural Networks (CNNs) have been used in the medical field for the detection, segmentation and classification of objects in medical images such as radiographs, Computed Tomography (CT), Magnetic Resonance Imaging (MRI) and endoscopic images [1]. Deep learning makes use of artificial neural networks to perform complicated computations on large amount of data (Website: Simplilearn). CNN minimises the need for manual feature extraction for image classification. Instead, the features are learned while the network processes the images [3]. CNN transforms the representation of data from lower levels into a more abstract high level using layers to make predictions.

Recently, Vision Transformers (ViTs) are developed and have achieved state of the art performance in Natural Language Processing. The success of the work achieved by [5] has ignited research in application of ViTs in image processing. For natural images, ViT has proved to outperform CNNs in standard computer vision work such as ImageNet classification, object detection as well as semantic segmentation. In comparison to convolutions, the attention mechanism at the heart of transformers has a number of significant advantages: (1) It depicts long-range relationships, (2) it has the ability to model adaptively using computed self-attention weights which allows capture of the relationship between tokens, (3) it offers a form of built-in saliency that provides information on what the model has focused on.

After performing an extensive literature review, it is found that only few works have been done to apply ViTs to endoscopic images, particularly to GI tract organs. Moreover, both CNNs and ViTs have been sparsely applied to the classification of colour misalignment in GI endoscopic images. In this work, we, therefore, apply both CNNs and ViTs to GI tract endoscopic images for the classification of colour misalignment artefacts and investigate their performance. The objective is to fill in the gap with regards to application of ViT to endoscopic images and comparison of the performance of ViT and CNN in the classification of endoscopic images corrupted with colour misalignment artefact. This will help to instigate further research in the application of ViT to endoscopic images, leading to investigation on the enhancement of ViT and CNN in the classification of colour misalignment artefacts as well as to artefact detection and classification in endoscopic images in general.

In the next section, we will, therefore, describe the work done so far using CNN and ViT. Then, the CNN and ViT models applied for the classification of colour misalignment are detailed. Finally, the results achieved are described and reviewed before proceeding with the conclusion.

2 LITERATURE REVIEW

Several works based on CNN have been undertaken for artefact detection and classification in endoscopic images. In [6], state of art deep learning techniques are investigated to detect and classify the precancerous stages of squamous cell carcinoma (SCC) cancer in real time during endoscopy. Image sequences corrupted by colour misalignment artefact including blur are eliminated first. Afterwards, the system allows the classification of the remaining esophageal video images into three classes of SCC: ‘suspicious’, ‘high grade’, and ‘cancer’. Conventional CNN, AlexNet is applied for the classification of colour misalignment. For detection and classification of images into the 3 SCC categories, Mask R-CNN and YOLOv3 are used. An accuracy of 96% is obtained for the artefact classification while for detection and classification of SCC, the accuracy results are 85% by YOLOV3 and 77% by Mask-R-CNN.

Multi-class artefact detection was performed by [24] for seven different artefact classes in endoscopic images including instrument, specularity, artefact, blur, contrast, bubbles, and saturation from the EAD 2019 dataset. An improved Cascade R-CNN model in combination with feature pyramid networks (FPN) is applied. The Cascade R-CNN achieves a good balance between mAP and IoU with a mAP of 0.3235 and IoU of 0.4172. [12] describes the deep learning architectures used in the EndoCV2020 challenge for the detection and segmentation of endoscopic artefacts and diseases in endoscopic images. A state-of-the-art detector, EfficientDet with different EfficientNet backbones and Focal length is trained and optimized for the detection task. The ensemble method provided the best detection performance with dScore of 0.44, a mean mAP of 0.36 and an IoU of 0.52.

[13] focuses on the detection of bounding boxes for seven classes of artefacts in the EAD 2019 dataset using Focal loss and RetinaNet architecture with Resnet-152 backbone. A 5-fold cross validation strategy has been used to optimize the parameters of the network. A mAP of 0.2719 has been obtained on 195 cases over 7 artefact classes. The IoU is 0.3456 for the detection task over the classes.

Transformers was introduced as a novel attention-driven technique for machine translation by [21]. [5] has proposed ViTs, based on the standard Transformer model of [21], by cascading a number of transformer layers to depict the global context of input images. Basically, an image is interpreted as a sequence of patches processed by a standard transformer encoder similar to the one used in NLP [15]. It has been concluded that when pre-trained on a wide set of data and transferred to medium-sized or smaller image classification tasks, ViT achieves quite promising results compared to CNN while consuming less computer resources during training [5]. The success of the ViT model has fueled the use of ViTs in the field of medical imaging, with applications in classification, object detection, and segmentation [15].

The authors in [10] investigated whether ViT models can replace CNNs in the field of medical image processing. In the process, several tests were carried out using 3 datasets: APTOS 2019 consisting of diabetic retinopathy images for classifying the latter into 5 disease severity categories, ISIC 2019 consisting of dermoscopic images depicting skin lesions for their classification into 9 classes of skin diseases and CBIS-DDSM consisting of mammography images where the task is the detection of masses in the images. The work concluded that when training is done from scratch on limited data, CNNs perform better than ViT due to the fact that the latter lacks inductive bias. When pre-trained on Imagenet, ViT performed comparably to CNNs with limited data. When transfer learning is used, both ViT and CNNs perform better. Most importantly, in case self-supervised training is applied followed by supervised fine-tuning, ViTs outperform CNNs in the field of medical diagnosis where data is limited. ResNet50 was used as the CNN based model and DeiT-S as the ViT while DINO was used as the self-supervised learning technique [10].

For the classification of ultrasound images with breast cancer into normal, malignant and benign, [8] evaluated the performance of ViT model against CNN. CNN models used are ResNet50, VGG16, Inception, and NASNET. Amongst the CNN based models, RestNet50 model achieved the best performance with an accuracy of 85.3% and Area Under Curve (AUC) of 0.95. ViT model using transfer learning attained an accuracy of 86.7% and AUC of 0.95. The results show the higher performance of the ViT model compared to that of CNN in terms of accuracy but for AUC both achieved similar results. ViTs has been used in the automated image classification in the context of Covid-19 diagnosis. [14] has used Point-of-Care Transformer (POCFormer) to classify Covid-19 from ultrasound images of the lungs and obtained above 91% average accuracy. POCFormer is a lightweight ViT which reduces the complexity of self-attention in terms of time and space from quadratic to linear. Following the analysis of previous work done, we find that CNN has been the main state of the art technique applied to endoscopic images and particularly to artefact detection. However, there is limited work done for the detection of colour misalignment artefact in endoscopic images using CNN. On the other hand, we find that there is a research gap in the application of ViT to endoscopic images. Therefore, we explore the application of both CNN and ViT in colour misalignment classification in endoscopic images and compare the performance.

3 METHODS

This section describes the experiment carried out to analyse the performance of ViT and CNN in the classification of colour misalignment artefact in endoscopic images.

3.1 Dataset

The data used in this study has been built using the dataset which has been released as part of Endoscopy Artefact Detection (EAD2019) IEEE ISBI'19 challenge [2]. It consists of 7 classes of artefacts namely saturations, motion blur, specularity, bubbles, imaging artefacts, contrast and instrument. The images with colour misalignment artefacts have been chosen manually. Of note, colour misalignment artefacts occur due to motion blur. Moreover, some videos containing colour misalignment artefacts have been collected from a previous study for SCC detection as detailed in [7]. The images were extracted from the videos using VideoProc Converter application [22]. In all, the number of images amount to 674 with 200 containing colour misalignment and 474 consisting of remaining artefacts. Hence, the dataset consists of 2 classes, one with colour misalignment and the second one with images of remaining artefacts. Part of the dataset is depicted in Figure 2.

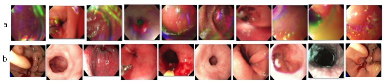

Figure 2: Part of the dataset (a - colour misalignment, b - other artefacts)

For both models implemented, one based on ViT and the other on a CNN architecture, the images have been divided into training, validation and test dataset in the ratio of 7, 2 and 1 respectively.

3.2 Model development

3.2.1 *ViT based model.*

Figure 3 shows the overall architecture of the method implemented for classification of colour misalignment artefact using the customised ViT model

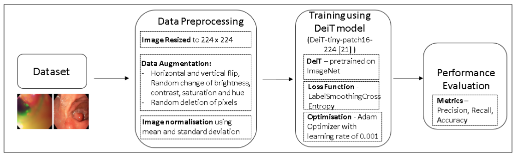
Dataset
Data Preprocessing
Image Resized to 224 x 224
Data Augmentation:
- Horizontal and vertical flip, Random change of brightness, contrast, saturation and hue
- Random deletion of pixels
Image normalisation using mean and standard deviation
Training using DeiT model
(DeiT-tiny-patch16-224 [21])
DeiT – pretrained on ImageNet
Loss Function - LabelSmoothingCross Entropy
Optimisation - Adam Optimizer with learning rate of 0.001
Performance Evaluation
Metrics – Precision, Recall, Accuracy

Figure 3: Model based on VIT

The images used for training have been augmented as part of the pre-processing step as shown in Figure 3 in order to have a better dataset without changing the meaning of the original images. The images in the validation and test dataset have been cropped and normalised.

The method implemented is based on the DeiT model - DeiT-tiny-patch16-224 [20] depicted in Figure 4.

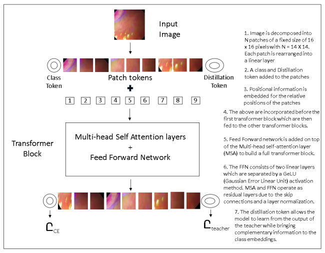
Input Image
Class Token
Patch tokens
Distillation Token
1 2 3 4 5 6 7 8 9
Transformer Block
Multi-head Self Attention layers
+
Feed Forward Network
ꞆCE
Ꞇteacher
1. Image is decomposed into N patches of a fixed size of 16 x 16 pixels with N = 14 X 14. Each patch is rearranged into a linear layer
2. A class and Distillation token added to the patches
3. Positional information is embedded for the relative positions of the patches
4. The above are incorporated before the first transformer block which are then fed to the other transformer blocks.
5. Feed Forward network is added on top of the Multi-head self-attention layer (MSA) to build a full transformer block.
6. The FFN consists of two linear layers which are separated by a GeLU (Gaussian Error Linear Unit) activation method. MSA and FFN operate as residual layers due to the skip connections and a layer normalization.
7. The distillation token allows the model to learn from the output of the teacher while bringing complementary information to the class embeddings.

Figure 4: DeIT architecture

DeiT is based on the ViT model developed by [5] which consider input images as a sequence of input tokens. The input images, which are of fixed sized and in RGB, are decomposed into N patches of a fixed size of 16 x 16 pixels with N = 14 X 14. Each patch is rearranged into a linear layer which conserves the overall dimension of the image (that is 3 X 16 X16). Positional information is embedded for the relative positions of the patches. These are incorporated before the first transformer block which are then fed to the other transformer blocks. A Feed Forward network is added on top of the

Multi-head self-attention layer (MSA) to build a full transformer block. The FFN consists of two linear layers which are separated by a GeLU activation method. MSA and FFN operate as residual layers due to the skip connections and a layer normalization. A class vector, which is a trainable vector, is appended to the patch tokens just before the first layer. The former goes through the transformer layers and is afterwards projected together with a linear layer in order to predict the class. A new token which is the distillation token is added to the patches and class tokens. It interacts using self-attention with the other embeddings and the network outputs it after the final layer. The purpose of the distillation token is to allow the model to learn from the output of the teacher while bringing complementary information to the class embeddings [20].

The DeiT was pre-trained on ImageNet database and transfer learning used to apply the algorithm to our customized dataset. LabelSmoothingCrossEntropy is used as the loss function while Adam optimizer with a learning rate of 0.001 has been used for optimisation.

3.2.2 CNN based model. The architecture of the model implemented based on CNN is as depicted in Figure 5. The dataset has been rescaled, horizontally flipped as well as sheared and zoomed as part of pre-processing. Four state of the art pre-trained CNN models have been used for experimentation. These are VGG16, ResNet50, Inceptionv3 and InceptionResNetV2.

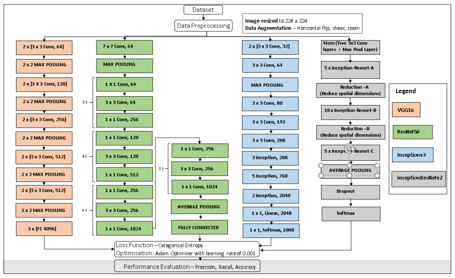

Figure 5: Architecture of CNN based model

VGG16 is a deep CNN which was introduced by [16]. It is made up of 16 layers with 13 convolutional layers and 3 fully connected layers. It also includes a final softmax classifier. 5 out of the 13 convolutional layers are max-pooling layers. ResNet50 was introduced by [9]. It is a CNN consisting of 50 layers with 48 convolutional layers, a maxpool layer and an average pool layer. ResNet-50 is formed by stacking residual blocks to form the network [4]. Inceptionv3 is a deep learning model which is based on the CNN developed in 2015 by a team at Google. It is an advanced version of the based model Inceptionv1 (GoogleNet). The network architecture is made up of 42 layers and has a better error rate than its predecessors. To achieve this, larger convolutions are factorized into smaller ones while making use of asymmetric convolutions. Auxiliary classifiers are used as regulariser and grid size is reduced efficiently [19]. InceptionResNetV2 is proposed by [18]. It is based on the Inception architecture. The stem module consists of two 3x3 convolutional layers followed by a max pooling layer while the Inception blocks make use of residual connections to overcome the degradation problem due to the deep structures. Reduction blocks are used to reduce the spatial dimensions. Overall the network consists of 164 layers and helps to reduce the training time.

In our implementation, categorical entropy is used as the loss function and similar to the ViT model, Adam algorithm with a learning rate of 0.001 is used as the optimization method. The dataset has been trained using each CNN model and their performance on the test dataset evaluated.

4 RESULTS AND DISCUSSION

Several experiments were carried out in order to determine the best classification model under different hyperparameters. The algorithms for both the ViT model and the CNN models were run at 25 and 50 epochs. For each number of epochs, the model was run 5 times and the average values of accuracy, precision and recall were calculated. Out of the 4 CNN models, the best validation accuracy of 92.79% was obtained at 50 epochs by InceptionResNetv2 followed by Inceptionv3 with a validation accuracy of 91.29%, VGG16 with 79.48% and ResNet50 with 76.12%. For the ViT implementation, the best validation accuracy of 96.27% was also obtained at 50 epochs. It can be concluded that better results are achieved when the number of epochs increases both for ViT and CNNs. Moreover, ViT provides better validation accuracy than CNN based models. These models were then run on the test data. We find that the ViT model based on the DeiT architecture (DeiT) achieved the best performance with an overall accuracy of 96.33% as compared to the Inceptionv3 with an accuracy of 78.67%, InceptionResNetv2 with 76.67%, VGG16 with 53.19% and finally ResNet50 with 51%. Additionally, the ViT model has the best precision and recall values on test data as compared to the CNN models. This could be explained with the ability of ViT to better depict the relationship between image pixels through self-attention weights. Figure 6. provides a graphical view of the accuracy, precision and recall results obtained with the different models. Overall ViT performs better than the CNN models when run on test data.

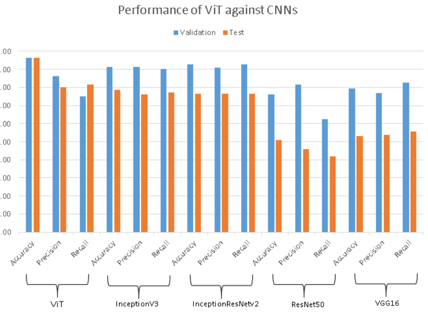

Figure 6: Performance of ViT against CNN

The confusion matrix obtained when the best ViT model was run on test data is as displayed in Figure 7

The matrix shows that the number of false positives and false negatives are quite low with the ViT model. Next, Figure 8 provides a visualization of the attention model when applied to endoscopic images with colour misalignment artefacts and those with the rest of artefacts.

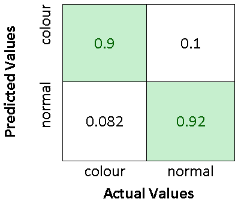

Figure 7: Confusion matrix with ViT model (colour – colour misalignment, normal – remaining artefacts)

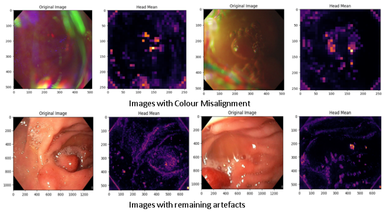

Figure 8: Visualizing attention in Vision Transformers

5 CONCLUSION

CNN based architectures have been the de facto techniques applied in the medical field for endoscopic image classification, object detection and segmentation. The application of ViT in medical imaging is emerging and has scope for further development. The main objective of this paper is to evaluate whether ViT can be applied efficiently to classification of colour misalignment artefacts in endoscopic images as compared to CNN based models. It can be concluded that ViT achieves better performance in terms of accuracy, precision and recall than CNN on the classification of colour misalignment artefacts from other artefacts namely saturations, specularity, bubbles, imaging artefacts, contrast and instrument in endoscopic images. Moreover, this study demonstrates that when applied to endoscopic images, ViT provides better performance than CNN based models. Furthermore, although datasets can be limited in size in the medical field, it is found that when pre-training is performed on large datasets and transfer learning applied, ViT can provide significant and promising results.

REFERENCES

[1] Sharib Ali, Mariia Dmitrieva, Noha Ghatwary, Sophia Bano, Gorkem Polat, Alptekin Temizel, Adrian Krenzer, Amar Hekalo, Yun Bo Guo, Bogdan Matuszewski, Mourad Gridach, Irina Voiculescu, Vishnusai Yoganand, Arnav Chavan, Aryan Raj, Nhan T. Nguyen, Dat Q. Tran, Le Duy Huynh, Nicolas Boutry, Shahadate Rezvy, Haijian Chen, Yoon Ho Choi, Anand Subramanian, Velmurugan Balasubramanian, Xiaohong W. Gao, Hongyu Hu, Yusheng Liao, Danail Stoyanov, Christian Daul, Stefano Realdon, Renato Cannizzaro, Dominique Lamarque, Terry Tran-Nguyen, Adam Bailey, Barbara Braden, James East, and Jens Rittscher. 2021. Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy. *Medical Image Analysis* 70 (May 2021), 102002. https://doi.org/10.1016/j.media.2021.102002 arXiv:2010.06034 [cs].

[2] Sharib Ali, Felix Zhou, Christian Daul, Barbara Braden, Adam Bailey, Stefano Realdon, James East, Georges Wagnières, Victor Loschenov, Enrico Grisan, Walter Blondel, and Jens Rittscher. 2019. Endoscopy artifact detection (EAD 2019) challenge dataset. https://doi.org/10.17632/C7FJBXCGJ9.1 arXiv:1905.03209 [cs, eess].

[3] Mayank Banoula. [n. d.]. What Is Deep Learning? | How It Works, Techniques & Applications.

[4] datagen.tech. [n. d.]. ResNet-50: The Basics and a Quick Tutorial.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://openreview.net/forum?id=YicbFdNTTy

[6] Xiaohong Gao, Barbara Braden, Stephen Taylor, and Wei Pang. 2019. Towards Real-Time Detection of Squamous Pre-Cancers from Oesophageal Endoscopic Videos. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. 1606–1612. https://doi.org/10.1109/ICMLA.2019.00264

[7] Xiaohong W. Gao, Stephen Taylor, Wei Pang, Rui Hui, Xin Lu, and Barbara Braden. 2023. Fusion of colour contrasted images for early detection of oesophageal squamous cell dysplasia from endoscopic videos in real time. *Information Fusion* 92 (April 2023), 64–79. https://doi.org/10.1016/j.inffus.2022.11.023

[8] Behnaz Gheflati and Hassan Rivaz. 2022. Vision Transformer for Classification of Breast Ultrasound Images. https://doi.org/10.48550/arXiv.2110.14731 arXiv:2110.14731 [cs].

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. https://doi.org/10.1109/CVPR.2016.90 ISSN: 1063-6919.

[10] Christos Matsoukas, Johan Haslum, Magnus Soderberg, and Kevin Smith. 2021. *Is it Time to Replace CNNs with Transformers for Medical Images?*

[11] Ken Namikawa, Toshiaki Hirasawa, Toshiyuki Yoshio, Junko Fujisaki, Tsuyoshi Ozawa, Soichiro Ishihara, Tomonori Aoki, Atsuo Yamada, Kazuhiko Koike, Hideo Suzuki, and Tomohiro Tada. 2020. Utilizing artificial intelligence in endoscopy: a clinician's guide. *Expert Review of Gastroenterology & Hepatology* 14, 8 (Aug. 2020), 689–706. https://doi.org/10.1080/17474124.2020.1779058

[12] Nhan T. Nguyen, Dat Q. Tran, and Dung B. Nguyen. 2020. Detection and Segmentation of Endoscopic Artefacts and Diseases Using Deep Architectures. https://doi.org/10.1101/2020.04.17.20070201 Pages: 2020.04.17.20070201.

[13] Ilkay Oksuz, James R. Clough, James R. Clough, and Julia A. Schnabel. 2019. Artefact detection in video endoscopy using retinanet and focal loss function. *CEUR Workshop Proceedings* 2366 (2019). http://www.scopus.com/inward/record.url?scp=85066467552&partnerID=8YFLogxK

[14] Shehan Perera, Srikar Adhikari, and Alper Yilmaz. 2021. POCFormer: A Lightweight Transformer Architecture for Detection of COVID-19 Using Point of Care Ultrasound. https://doi.org/10.48550/arXiv.2105.09913 arXiv:2105.09913 [cs, eess].

[15] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. 2022. Transformers in Medical Imaging: A Survey. https://doi.org/10.48550/arXiv.2201.09873 arXiv:2201.09873 [cs, eess].

[16] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. https://doi.org/10.48550/arXiv.1409.1556 arXiv:1409.1556 [cs].

[17] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. 2021. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: a cancer journal for clinicians* 71, 3 (May 2021), 209–249. https://doi.org/10.3322/caac.21660

[18] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. 2016. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. https://doi.org/10.48550/arXiv.1602.07261 arXiv:1602.07261 [cs].

[19] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the Inception Architecture for Computer Vision. https://doi.org/10.48550/arXiv.1512.00567 arXiv:1512.00567 [cs].

[20] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. https://doi.org/10.48550/arXiv.2012.12877 arXiv:2012.12877 [cs].

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[22] VideoProc. [n. d.]. [OFFICIAL] VideoProc Converter - One-Stop Video Processing Software for Windows Mac. https://www.videoproc.com/

[23] Lianlian Wu, Wei Zhou, Xinyue Wan, Jun Zhang, Lei Shen, Shan Hu, Qianshan Ding, Ganggang Mu, Anning Yin, Xu Huang, Jun Liu, Xiaoda Jiang, Zhengqiang Wang, Yunchao Deng, Mei Liu, Rong Lin, Tingsheng Ling, Peng Li, Qi Wu, Peng Jin, Jie Chen, and Honggang Yu. 2019. A deep neural network improves endoscopic detection of early gastric cancer without blind spots. *Endoscopy* 51, 6 (June 2019), 522–531. https://doi.org/10.1055/a-0855-3532

[24] Suhui Yang and G. Cheng. 2019. ENDOSCOPIC ARTEFACT DETECTION AND SEGMENTATION WITH DEEP CONVOLUTIONAL NEURAL NETWORK. https://www.semanticscholar.org/paper/ENDOSCOPIC-ARTEFACT-DETECTION-AND-SEGMENTATION-WITH-Yang-Cheng/57c589a70e3dd1b9fcb57ccd7361387ddfc3e8ed