CNNs vs. Transformers: Performance and Robustness in Endoscopic Image Analysis

Carolus H. J. Kusters[1(✉)], Tim G. W. Boers[1], Tim J. M. Jaspers[1], Jelmer B. Jukema[2], Martijn R. Jong[2], Kiki N. Fockens[2], Albert J. de Groof[2], Jacques J. Bergman[2], Fons van der Sommen[1], and Peter H. N. de With[1]

[1] Department of Electrical Engineering, Video Coding and Architectures, Eindhoven University of Technology, Eindhoven, The Netherlands
{c.h.j.kusters,t.boers,t.j.m.jaspers,fvdsommen,p.h.n.de.with}@tue.nl
[2] Department of Gastroenterology and Hepatology, Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, The Netherlands

Abstract. In endoscopy, imaging conditions are often challenging due to organ movement, user dependence, fluctuations in video quality and real-time processing, which pose requirements on the performance, robustness and complexity of computer-based analysis techniques. This paper poses the question whether Transformer-based architectures, which are capable to directly capture global contextual information, can handle the aforementioned endoscopic conditions and even outperform the established Convolutional Neural Networks (CNNs) for this task. To this end, we evaluate and compare clinically relevant performance and robustness of CNNs and Transformers for neoplasia detection in Barrett's esophagus. We have selected several top performing CNN and Transformers on endoscopic benchmarks, which we have trained and validated on a total of 10,208 images (2,079 patients), and tested on a total of 4,661 images (743 patients), divided over a high-quality test set and three different robustness test sets. Our results show that Transformers generally perform better on classification and segmentation for the high-quality challenging test set, and show on-par or increased robustness to various clinically relevant input data variations, while requiring comparable model complexity. This robustness against challenging video-related conditions and equipment variations over the hospitals is an essential trait for adoption in clinical practice. The code is made publicly available at: https://github.com/BONS-AI-VCA-AMC/Endoscopy-CNNs-vs-Transformers.

Keywords: Barrett's Esophagus · CNN · Transformers · Robustness

C. H. J. Kusters—This work is facilitated by data/equipment from Olympus Corp., Tokyo, Japan.

S. Wu et al. (Eds.): AMAI 2023, LNCS 14313, pp. 21–31, 2024.
https://doi.org/10.1007/978-3-031-47076-9_3

1 Introduction

Due to the increase in image quality of endoscopy devices, endoscopic image analysis by visual inspection has become an essential aspect in gastrointestinal endoscopy. The use of endoscopic procedures allows for non-invasive diagnosis and treatment of various gastrointestinal complications, such as cancer, ulcers and inflammatory diseases. However, the analysis of endoscopic imagery can be challenging due to operator dependence, low inter-observer agreement, variability in image quality, artifacts and subtle differences between normal and abnormal tissue. To address these aspects, CAD techniques have been developed to aid endoscopists in the diagnosis and treatment of gastrointestinal complications.

A recent technique that has emerged as a promising solution is deep learning-based AI. AI-based tools have the potential to attractively benefit the diagnostic accuracy and efficiency of endoscopic image analysis by endoscopists. Several AI-based tools have been proposed for the detection and classification of suspected neoplastic lesions in Barrett's esophagus (BE) [9–11,14], colorectal polyps [3,6] and gastric lesions [7,23]. For the aforementioned applications, Convolutional Neural Networks (CNNs) are considered the state-of-the-art solution.

Recently, the Vision Transformer (ViT) [8] adopted the Transformer [21] architecture originally developed for NLP tasks, in the field of Computer Vision (CV). The ViT and its derived instances [15,22,25] are able to directly capture long-range feature dependencies with the self-attention mechanism, unlike CNNs which usually have a limited receptive field. Transformers have quickly gained popularity in a wide variety of CV problems and are competing with CNNs for state-of-the-art performance in many CV tasks and applications.

The use of CNNs in endoscopic image analysis has been the state-of-the-art for several years. However, the increasing need for robustness in light of the posed challenges and the benefits of Transformers, raises the question whether Transformer-based networks are more suitable in this field. Therefore, in this paper, we evaluate and compare performance and robustness of CNNs and Transformers for endoscopic image analysis, particularly, neoplasia detection in BE. This study uses one high-quality test set, enriched with challenging subtle neoplasia cases, to evaluate clinically relevant performance. Furthermore, we selected three different robustness test sets to evaluate clinically relevant robustness against low-quality data and/or out-of-domain data. Additionally, the memory requirements of the architectures are assessed to determine their suitability for deployment in clinical practice.

This study is the first in its kind and aims to provide insights into the comparative efficacy of CNNs and Transformers for endoscopic image analysis, in terms of the performance, robustness and complexity in light of the posed challenges.

2 Methods

2.1 Data: Setting, Datasets and Preprocessing

A. Setting: This study uses private internal data collected both retrospectively and prospectively at 15 international centers using Olympus gastroscopes

(Olympus Corp., Tokyo, Japan), comprising of Barrett's neoplasia and non-dysplastic BE (NDBE) imagery (with histopathology confirmation), and curated for presence (neoplasia) or absence (NDBE) of a visible lesion, respectively.

B. Datasets: *1. Training, Validation and Test Sets:* The split between train/validation/test datasets is made in accordance with clinicians, to ensure representative training and validation sets, while the test set is enriched with more challenging subtle neoplasia cases. An example of a subtle neoplasia case from the test set can be observed in Fig. 1a. A strict split on patient basis is employed to avoid data leakage and intra-patient bias. Delineations of neoplasia are obtained from 14 Barrett's experts, where at least 2 experts delineated the same image. For each image, the experts delineated the largest area that is suspected to be neoplasia (LL) and the area within the LL that stands out more profoundly (HL). To achieve a minimal level of consensus, a third expert endoscopist is invited in case the two HL delineations obtained less than 30% agreement in terms of Dice Score. Subsequently, the two delineations among the three experts that achieved the highest overlap are used for further ground-truth (GT) processing. For training and evaluation purposes, a consensus GT mask is constructed by means of (1) union of HL delineations (2) intersection of LL delineations, and (3) the union of (1) and (2). A detailed summary of the datasets, in terms of images, patients and GT masks is presented in Table 1.

2. Internal Robustness Test Sets: For internal robustness evaluation, two different test sets are constructed from imagery that is excluded for training and validation. The image-quality robustness test set (QRT) consists of images that are excluded due to inferior subjective image quality, and can be used to evaluate robustness against subjective image quality. The quality exclusion criteria are defined as the distance to the lesion, illumination, blur, contraction of the esophagus, resolution of imagery and presence of mucus and bubbles. The image-criteria robustness test set (CRT) consists of images that are excluded based on the presence or absence of a visible lesion for NDBE and neoplasia, respectively, and can be used to evaluate robustness and generalization on data not matching the inclusion criteria of the training set. All patients included in QRT and CRT sets are excluded from the algorithm training and validation sets to avoid data-leakage and intra-patient bias and no GT masks are available. Example imagery of the QRT/CRT sets can be observed in Fig. 1c and 1d, while a summary of the data statistics is presented in Table 1.

3. External Robustness Test Set: For external robustness evaluation, a test set (BORN) is constructed from frames extracted from the BORN training module videos [1]. The data in this module was retrospectively collected at 3 international centers, with older generations Olympus (Olympus Corp., Tokyo, Japan) gastroscopes, compared to internal data collection, and Fujifilm (Fujifilm Corp., Japan) gastroscopes. The BORN set can be used to evaluate robustness against objective image quality, scope type and scope manufacturer. An example of a low-quality image of different scope manufacturer can be observed in Fig. 1b. Neoplasia delineations are obtained from 4 Barrett's experts, where at least

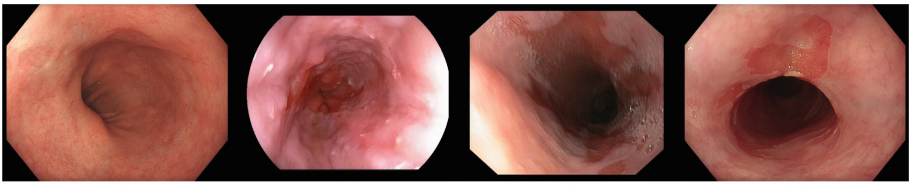


Fig. 1. (a) Test: high-quality subtle neoplasia case (b) BORN: low-quality image of different scope manufacturer (c) QRT: low-quality image (blur and presence of bubbles) (d) CRT: NDBE case with visible abnormality

3 experts delineated the same image. For evaluation purposes, a consensus GT mask is constructed by means of the intersection of the three individual delineations. A summary of data statistics is presented in Table 1.

Table 1. Description of datasets used for algorithm development and evaluation.

Dataset	NDBE	Neoplasia	GT Masks
Train	5,566 (948 pt)	4,442 (1,038 pt)	1,947
Validation	100 (36 pt)	100 (58 pt)	96
Test	300 (125 pt)	100 (50 pt)	100
QRT	109 (84 pt)	463 (248 pt)	/
CRT	150 (97 pt)	637 (93 pt)	/
BORN	1,601 (32 pt)	1,301 (65 pt)	1,301

C. Data Pre-processing: The central active region of raw endoscopic images is resized to 256×256 pixels, prior to normalizing the intensity values by channel-wise subtracting the mean and dividing by the standard deviation of the training data. Data augmentation techniques are applied during training to virtually increase the set size and to improve generalization. A random combination is employed of horizontal and vertical flipping, rotation by $\theta \in \{0°, 90°, 180°, 270°\}$, contrast/saturation/brightness enhancements, gray-scale conversion, Gaussian blurring, random affine and sharpness transforms, followed by random artificial Gaussian noise corruption. To obtain a reliable performance estimate of the model, validation evaluation is performed four times on a different randomly augmented internal validation set, by employing the first three augmentation techniques, to increase the sample size by a factor of four.

2.2 Network Architectures, Training and Evaluation

A. Network Architectures: In this study, four CNNs and three Transformer-based architectures are compared. The CNN selection includes the well-known

Table 2. Comparison of architectures used in this study, in terms of parameters, model size, multiply-add operations (M-A) and processing speed.

Network (Backbone)	Param.	Mod. Size	M-A	Proc. speed
U-Net (ResNet-50)	32.5 M	130.1 MB	127.6 G	5.95 fps
U-Net (ResNet-152)	67.2 M	268.6 MB	244.0 G	3.23 fps
U-Net++ (ResNet-50)	48.9 M	196.0 MB	689.3 G	2.39 fps
U-Net++ (ResNet-152)	83.6 M	334.5 MB	805.7 G	1.78 fps
DeepLabV3+ (ResNet-50)	40.3 M	161.4 MB	139.6 G	5.96 fps
DeepLabV3+ (ResNet-152)	74.9 M	299.9 MB	256.0 G	3.08 fps
DeepLabV3+ (ConvNeXt-T)	35.2 M	134.4 MB	75.8 G	7.55 fps
DeepLabV3+ (ConvNeXt-B)	96.8 M	387.1 MB	79.4 G	2.96 fps
CaraNet (Res2Net-101)	44.6 M	172.7 MB	143.4 G	4.56 fps
ESFPNet-T (MiT-B0)	3.5 M	14.01 MB	2.22 G	22.97 fps
ESFPNet-L (MiT-B4)	61.7 M	245.5 MB	20.6 G	2.51 fps
FCBFormer-B0 (PVTv2-B0)	11.3 M	45.3 MB	386.3 G	1.65 fps
FCBFormer-B3 (PVTv2-B3)	52.9 M	211.2 MB	536.6 G	1.16 fps
Swin-T-UperNet (SwinV2-T)	40.8 M	137.4 MB	30.9 G	6.44 fps
Swin-B-UperNet (SwinV2-B)	110.4 M	357.9 MB	55.3 G	2.45 fps

U-Net [18], U-Net++ [26] and DeepLabV3+ [5] architectures, with ResNet [12] and ConvNeXt [16] backbones. The more advanced CaraNet [17] is also included, which is based on context axial reverse attention. The Transformer-based architecture selection includes ESFPNet [4], FCBFormer [19] and UperNet [24], with Mix Transformer (MiT) [25], Pyramid Vision Transformer v2 (PVTv2) [22] and Swin Transformer V2 (SwinV2) [15] backbones, respectively. All backbones are initialized with ImageNet-pretrained weights, and extended with a simple classification head consisting of a downsampling stage and a single fully-connected layer with a Sigmoid activation function, after the final backbone feature extraction layer. Based on the endoscopic segmentation benchmarks (Kvasir-SEG [13], CVC-ColonDB [20] and CVC-ClinicDB [2]) performance, CaraNet, ESFPNet and FCBFormer are selected for comparison, while the other architectures are applied in generic segmentation problems. The original source codes of the architectures are adapted to fit our use case. A comparison of the architectures in terms of parameters, model size, multiply-add operations (M-A) and processing speed (measured on 12-Core Ryzen 9 5900X CPU) is listed in Table 2.

B. Training Procedures: All architectures are trained in a two-step procedure with batch sizes of 32 (16 for FCBFormer-B3) and learning rates in the range of 10^{-3} - 10^{-7}, optimized by a maximum of three runs for each architecture, after which the best model is used for a single-evaluation iteration on the test sets. First, training with all available data for 150 epochs is performed, followed by training with prospectively collected data for 75 epochs. Early stopping is

Table 3. Performance on all test sets for all architectures

	Test Set		BORN Set		QRT	CRT
Network (Backbone)	AUC_{cls}	mD_i	AUC_{cls}	mD_i	AUC_{cls}	AUC_{cls}
U-Net (ResNet-50)	0.867	0.353	0.861	0.271	0.829	0.820
U-Net (ResNet-152)	0.938	0.529	0.849	0.313	0.878	0.782
U-Net++ (ResNet-50)	0.904	0.440	0.878	0.323	0.878	0.813
U-Net++ (ResNet-152)	0.898	0.451	0.860	0.308	0.868	0.780
DeepLabV3+ (ResNet-50)	0.860	0.410	0.825	0.308	0.799	0.784
DeepLabV3+ (ResNet-152)	0.926	0.456	0.885	0.296	0.854	0.789
DeepLabV3+ (ConvNeXt-T)	0.939	0.440	0.838	0.283	0.890	0.838
DeepLabV3+ (ConvNeXt-B)	0.909	0.453	0.862	0.291	0.871	0.820
CaraNet (Res2Net-101)	0.907	0.397	**0.897**	**0.344**	0.873	0.797
ESFPNet-T (MiT-B0)	0.908	0.424	0.850	0.209	0.872	0.806
ESFPNet-L (MiT-B4)	0.941	0.500	0.865	0.311	0.891	0.834
FCBFormer-B0 (PVTv2-B0)	0.896	0.470	0.861	0.283	0.877	0.828
FCBFormer-B3 (PVTv2-B3)	**0.961**	**0.552**	0.889	0.338	0.904	0.816
Swin-T-UperNet (SwinV2-T)	0.919	0.534	0.872	0.303	**0.910**	**0.842**
Swin-B-UperNet (SwinV2-B)	0.925	0.491	0.884	0.294	0.895	0.835

applied when the loss on the validation set has converged. The Adam optimizer with AMS-grad is used with (β_1, β_2)=(0.9, 0.999) and a weight decay of 10^{-4}, in combination with a learning rate scheduler that reduced the learning rate with a factor 10, when the validation loss stopped improving for 10 epochs with a maximum of 3 reductions. The classification loss of the algorithm is evaluated using the Binary Cross-Entropy (BCE) loss function, while the segmentation loss is evaluated with a compound BCE + Dice loss function, both with label smoothing of 0.01. To efficiently leverage images without a GT delineation, only the classification loss is used for backpropagation to improve training of the backbone encoder. Randomly sampling training images ensures an average 50-50% representation of classes in each iteration. Experiments are implemented in Python 3.10 using the PyTorch (Lightning) frameworks.

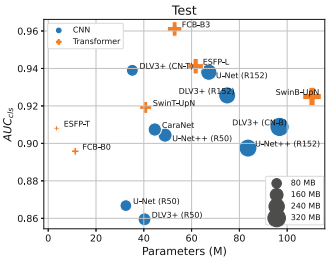
C. Performance Evaluation Metrics: The Area under the Curve (AUC) for the receiver operating characteristic (ROC), is used for evaluation of the classification (AUC_{cls}) performance. The metric is computed based on the classification branch output of the networks. The segmentation performance for neoplastic imagery is evaluated with the mean Dice Score (mD_i), by employing the segmentation mask, thresholded by 0.5, and the GT delineation.

3 Experimental Results and Discussion

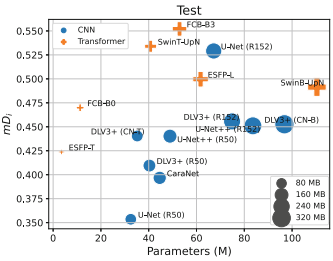
The results on all test sets are presented in Table 3, while graphical illustrations of the comparison between the size requirements (Parameters and Model Size) and the performance metrics for each set are depicted in Fig. 2.

1. General Performance on Test Set: The results indicate that FCBFormer-B3 outperforms other architectures on both classification and segmentation, with improvements of 0.02 and 0.018, respectively. As shown in Fig. 2a and 2b, the best performing classification and segmentation models are in the middle segment of parameter and model size, while the models in the low and high segment achieve similar performance. This suggests that having more parameters does not necessarily result in improved performance. Furthermore, a key observation is that Transformers achieve the top-2 scores for both classification and segmentation, and are well-represented in the top-5. These results suggest that Transformers generally perform better than CNNs on the classification and segmentation of challenging clinical cases of BE neoplasia, while requiring a comparable complexity in terms of parameters and model size.

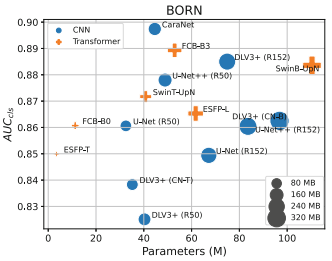
2. Robustness on BORN, QRT and CRT sets: The results on the BORN set indicate that CaraNet outperforms other architectures for both the classification and segmentation tasks, with margins of 0.008 and 0.006, respectively. As seen in Fig. 2c and 2d, the top classification and segmentation models are in the middle segment of parameter and model size, while models in the low and high segments achieve similar performance. This reinforces the conviction that having more parameters does not necessarily induce an improved performance. An important observation is that Transformers are well-represented in the top-5 of scores for both classification and segmentation. These findings indicate that Transformers achieve on-par robustness and generalization performance with CNNs on classification and segmentation of BE neoplasia in clinically realistic lower-quality data, obtained from older generation and different manufacturers' gastroscopes, while requiring a comparable model complexity. On the QRT and CRT sets, Swin-T-UperNet has the highest classification performance with margins of 0.006 and 0.004, respectively. As seen in Fig. 2e and 2f, the best performing models on the QRT set are found in the middle and high segments, which suggests that, in this case, having more parameters *is indeed* beneficial for performance, while the best performing models on the CRT set are found in all segments. It is noteworthy, that Transformers overshadow CNNs in the top-5 of scores for both the QRT and CRT sets, which indicates that Transformers show increased robustness compared to CNNs in clinically realistic data of lower subjective image quality or data that is not meeting the inclusion criteria of the training set, while requiring similar or less model complexity.



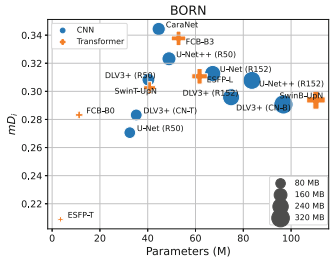
(a) Test: Size requirements vs. AUC_{cls}



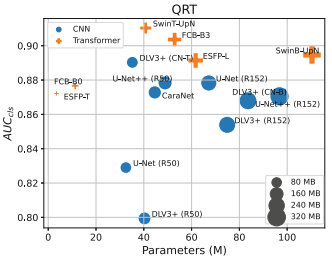
(b) Test: Size requirements vs. mD_i



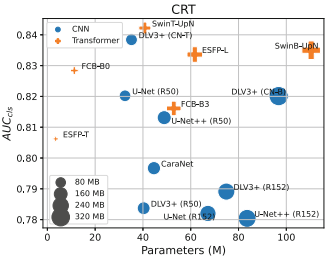
(c) BORN: Size requirements vs. AUC_{cls}



(d) BORN: Size requirements vs. mD_i



(e) QRT: Size requirements vs. AUC_{cls}



(f) CRT: Size requirements vs. AUC_{cls}

Fig. 2. Size requirements of all architectures, in terms of parameters and model size, versus the performance metrics on the Test, BORN, QRT and CRT sets.

4 Conclusions

Endoscopic image analysis is subject to several challenges, posing requirements on the performance, robustness and complexity of computer-based analysis techniques. These requirements and the benefits of the recently proposed Transformer-based architectures, raise the question whether Transformers are more suitable than state-of-the-art CNNs in this field. In this paper, we have evaluated and compared the performance, robustness and complexity of CNNs and Transformers for endoscopic image analysis, in this case Barrett's neoplasia detection. To this end, we have selected one high-quality test set and three different robustness test sets. The results show that Transformers generally outperform CNNs by a small margin on both classification and segmentation of

challenging subtle clinical representative Barrett's neoplasia cases. Additionally, they exhibit on-par or increased robustness various to clinically realistic and relevant data variations, while requiring comparable complexity in terms of parameters and model size, for this specific application. These findings are overall promising for the deployment of Transformers into endoscopy clinical practice, as they show critical robustness and generalization traits in case of varying quality and equipment over hospitals. However, future research is required to generalize and strengthen these findings, which should focus on (1) extension to other endoscopic applications, (2) evaluating additional aspects, such as efficiency in training sample size, robustness to clinically realistic artificial image corruption and the comparison with expert endoscopist performance.

Prospect of Application: Providing valuable insights into the comparative efficacy of CNNs and Transformers in endoscopic image analysis, addressing domain-specific challenges and emphasizing essential traits such as robustness, generalization and complexity, crucial for reliable operation in the diverse and challenging imaging nature of endoscopy clinical practice. Future research is needed to generalize and reinforce the findings, ensuring applicability beyond the current scope.

References

1. Bergman, J.J., de Groof, A.J., et al.: An interactive web-based educational tool improves detection and delineation of Barrett's esophagus-related neoplasia. Gastroenterol. **156**(5), 1299-1308.e3 (2019). https://doi.org/10.1053/j.gastro.2018.12.021
2. Bernal, J., et al.: WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. Comput. Med. Imaging Graph. **43**, 99–111 (2015). https://doi.org/10.1016/j.compmedimag.2015.02.007
3. Byrne, M.F., et al.: Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. Gut **68**(1), 94–100 (2019)
4. Chang, Q., et al.: ESFPNet: efficient deep learning architecture for real-time lesion segmentation in autofluorescence bronchoscopic video. In: Gimi, B.S., Krol, A. (eds.) Medical Imaging 2023: Biomedical Applications in Molecular, Structural, and Functional Imaging, vol. 12468, p. 1246803. International Society for Optics and Photonics, SPIE (2023). https://doi.org/10.1117/12.2647897
5. Chen, L.C., et al.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 833–851. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_49
6. Chen, P.J., et al.: Accurate classification of diminutive colorectal polyps using computer-aided analysis. Gastroenterol. **154**(3), 568–575 (2018)
7. Cho, B.J., et al.: Automated classification of gastric neoplasms in endoscopic images using a convolutional neural network. Endosc. **51**(12), 1121–1129 (2019)
8. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. ICLR (2021)

9. Ebigbo, A., et al.: Real-time use of artificial intelligence in the evaluation of cancer in barrett's oesophagus. Gut **69**(4), 615–616 (2020)
10. de Groof, A.J., et al.: Deep-learning system detects neoplasia in patients with Barrett's esophagus with higher accuracy than endoscopists in a multistep training and validation study with benchmarking. Gastroenterol. **158**(4), 915–929 (2020)
11. Hashimoto, R., et al.: Artificial intelligence using convolutional neural networks for real-time detection of early esophageal neoplasia in Barrett's esophagus (with video). Gastrointest. Endosc. **91**(6), 1264–1271 (2020)
12. He, K., et al.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
13. Jha, D., et al.: Kvasir-SEG: a segmented polyp dataset. In: Ro, Y.M., et al. (eds.) MMM 2020. LNCS, vol. 11962, pp. 451–462. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37734-2_37
14. Kusters, C.H.J., et al.: A CAD system for real-time characterization of neoplasia in Barrett's esophagus NBI videos. In: Ali, S., van der Sommen, F., Papież, B.W., van Eijnatten, M., Jin, Y., Kolenbrander, I. (eds.) Cancer Prevention Through Early Detection, pp. 89–98. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-17979-2_9
15. Liu, Z., et al.: Swin Transformer V2: scaling up capacity and resolution. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11999–12009 (2022). https://doi.org/10.1109/CVPR52688.2022.01170
16. Liu, Z., et al.: A convnet for the 2020s. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11966–11976 (2022). https://doi.org/10.1109/CVPR52688.2022.01167
17. Lou, A., et al.: CaraNet: context axial reverse attention network for segmentation of small medical objects. In: Medical Imaging 2022: Image Processing, vol. 12032, pp. 81–92. International Society for Optics and Photonics, SPIE (2022). https://doi.org/10.1117/12.2611802
18. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
19. Sanderson, E., Matuszewski, B.J.: FCN-transformer feature fusion for polyp segmentation. In: Yang, G., Aviles-Rivero, A., Roberts, M., Schönlieb, C.B. (eds.) Medical Image Understanding and Analysis, pp. 892–907. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-12053-4_65
20. Tajbakhsh, N., et al.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE Trans. Med. Imaging **35**(2), 630–644 (2016). https://doi.org/10.1109/TMI.2015.2487997
21. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017)
22. Wang, W., et al.: PVT v2: improved baselines with pyramid vision transformer. Comput. Vis. Media , 1–10 (2022). https://doi.org/10.1007/s41095-022-0274-8
23. Wu, L., et al.: Deep learning system compared with expert endoscopists in predicting early gastric cancer and its invasion depth and differentiation status (with videos). Gastrointest. Endosc. **95**(1), 92–104.e3 (2022). https://doi.org/10.1016/j.gie.2021.06.033

24. Xiao, T., et al.: Unified perceptual parsing for scene understanding. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11209, pp. 432–448. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01228-1_26
25. Xie, E., et al.: SegFormer: simple and efficient design for semantic segmentation with transformers. In: Neural Information Processing Systems (NeurIPS) (2021)
26. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: a nested u-net architecture for medical image segmentation. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 3–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_1