A New Dataset and Baseline Model for Rectal Cancer Risk Assessment in Endoscopic Ultrasound Videos

Jiansong Zhang[1,2(✉)], Shengnan Wu[3], Peizhong Liu[2], and Linlin Shen[1,4,5(✉)]

[1] School of Computer Science & Software Engineering, Shenzhen University, Shenzhen, China
[2] School of Medicine, Huaqiao University, Quanzhou, China
[3] Department of Medical Ultrasound, The First Affiliated Hospital of Fujian Medical University, Fuzhou, China
[4] AI Research Center for Medical Image Analysis and Diagnosis, Shenzhen University, Shenzhen, China
[5] Guangdong Provincial Key Laboratory, Shenzhen, Shenzhen, China
✉ Corresponding: jasonzhang3111@gmail.com, llshen@szu.edu.cn

Abstract. Early diagnosis of rectal cancer is essential to improve patient survival. Existing diagnostic methods mainly rely on complex MRI as well as pathology-level co-diagnosis. In contrast, in this paper, we collect and annotate for the first time a rectal cancer ultrasound endoscopy video dataset containing 207 patients for rectal cancer video risk assessment. Additionally, we introduce the Rectal Cancer Video Risk Assessment Network (RCVA-Net), a temporal logic-based framework designed to tackle the classification of rectal cancer ultrasound endoscopy videos. In RCVA-Net, we propose a novel adjacent frames fusion module that effectively integrates the temporal local features from the original video with the global features of the sampled video frames. The intra-video fusion module is employed to capture and learn the temporal dynamics between neighbouring video frames, enhancing the network's ability to discern subtle nuances in video sequences. Furthermore, we enhance the classification of rectal cancer by randomly incorporating video-level features extracted from the original videos, thereby significantly boosting the performance of rectal cancer classification using ultrasound endoscopic videos. Experimental results on our labelled dataset show that our RCVA-Net can serve as a scalable baseline model with leading performance. The code of this paper can be accessed at:https://github.com/JsongZhang/RCVA-Net

Keywords: Rectal cancer · Ultrasound endoscopy video dataset · Ultrasound video classification.

1 Introduction

Rectal cancer is one of the malignant tumours with the highest morbidity and mortality rates worldwide[5]. Due to its insidious onset, it is often difficult to

be noticed at first, which leads to a poor prognosis for patients. The existing diagnosis of rectal cancer is based on the combination of MRI[15,11] and pathological screening[7]. In recent years, ultrasound has attracted much attention as a simple and convenient imaging method[1,10,23]. With the design of the probe, doctors are now able to use endoscopic ultrasound to analyse rectal cancer via the anus. Endoscopic ultrasound imaging is intuitive and effective for the diagnosis of rectal cancer[16]. It significantly reduces diagnostic costs compared to the aforementioned co-diagnostics and is superior to simple superficial endoscopic examinations due to the collection of acoustic imaging in the inner periphery[22]. Moreover, the dynamic endoscopic ultrasound video visualises the ultrasonographic features of different cancerous conditions compared to static images. However, current research predominantly focuses on statistical analysis of static images[9,19], while deep learning studies centred around video modalities receive comparatively less attention[24].

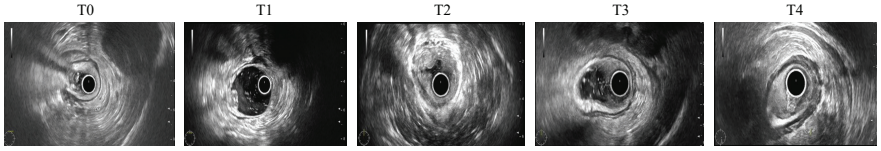

Fig. 1. Illustration of the five endoscopic ultrasound categories of rectal cancer in RCEUV-207, graded T0-T4 malignancy from left to right.

For this reason, this paper first presents an endoscopic ultrasound video dataset (RCEUV-207) of rectal cancer containing five categories to enhance the research in this field, with different categories of lesion intercepts as shown in Figure 1. To the best of our knowledge, this is the first benchmark dataset in the field. Existing ultrasound analyses are almost always based on static images. Whereas, for variable ultrasound sweeps, video data would not only provide information in the temporal dimension but also be more relevant. In addition, we performed extensive benchmarking on this dataset and based on this, we proposed a temporal fusion-based endoscopic ultrasound risk assessment network for rectal cancer (RCVA-Net). The contributions of this paper can be summarised as follows:

- We present the rectal cancer endoscopic ultrasound video dataset (RCEUV-207) and detail the challenges encountered in ultrasound video analysis of rectal cancer. This dataset enhances research in the field of computer-aided diagnosis for rectal cancer, focusing on ultrasound video insights.
- We conducted comprehensive benchmarking at the image and video levels for RCEUV-207 and proposed RCVA-Net, delivering key insights for this field-specific analysis.
- Experimental results demonstrate that RCEUV-207 harbours significant research potential. Our scalable RCVA-Net model exhibits leading perfor-

mance in this field, setting new benchmarks for accuracy and efficiency in this field.

2 The Proposed RCEUV Dataset

This paper presents ultrasound endoscopy data on rectal cancer collected in the First Hospital of Fujian Medical University, which consists of 207 independent patients collected from 2019-2023 (Ethical Review No. 23122). In RCEUV-207, there are five categories corresponding to the five stages of rectal cancer malignancy pathology T0-T4. All category labels of the video data have strict pathological diagnoses corresponding to them, which ensures that endoscopic ultrasound results for rectal cancer based on ultrasound videos are trustworthy.

2.1 Data Statistics

Rectal endoscopic ultrasound is an imaging technique in which a miniature ultrasound probe is delivered through the anus into the rectum. Not only does it provide a clear picture of the lesions contained within the rectal lining, but due to the ability of ultrasound to penetrate the intestinal tissues, transrectal ultrasound can also gather invasive information about the diseased tissue. In RCEUV-207, we used this technique to collect data from a population of patients with five different grades of rectal cancer stages. In the RCEUV-207, T0 indicates cases with completely non-cancerous tissues or completely regressed carcinomas, totalling 32 cases. T1 indicates cases where the carcinoma invades the submucosa, totalling 14 cases. T2 indicates cases of carcinoma invasion into the muscular layer, totalling 72 cases. T3 indicates tumours that invade through the muscular layer, totalling 81 cases. T4 indicates cases where the carcinoma invades other organs or structures, even through the peritoneum, with such patients being less common, totalling only 8 cases. It is worth noting that we performed one process on all raw video acquisitions. Except for the video data of T0 staging, all other categories of data have been processed based on the original collection. This indicates that the diseased tissue will not appear directly at the beginning of the video, but will be present in the middle of the video. We used this form of data to realistically simulate what would occur in the application. In addition, to ensure full exposure of the malignant tissue in the ultrasound view, the length of the videos included in RCEUV-207 varies from 15 to 60s.

2.2 Challenges in RCEUV

Rectal cancer ultrasound diagnosis based on endoscopic ultrasound (EUS) means faces several challenges when analysed using deep learning techniques. Although deep learning has shown great potential in the field of medical image processing[14], its application to EUS rectal cancer diagnosis specifically presents the following challenges:

Correlation of labels and data: When applying deep learning to medical image analysis in ultrasound diagnosis of rectal cancer based on endoscopic ultrasound means, the correlation between labels and data is a core element in achieving a highly accurate model. This relates to the accuracy and consistency of data labelling and how to ensure that the labels truly reflect the medical information in the images. Since endoscopic ultrasound is often not used as a direct basis for staging diagnosis of rectal cancer at this stage of clinical practice, mining effective data processing paradigms has become a challenge in this area. The establishment of an endoscopic ultrasound-based risk assessment network for rectal cancer grading based on effective pattern recognition tools to achieve effective feature extraction and risk rating is an important task in the RCEUV-207 analysis. In RCEUV-207, we ensured that each video category label was derived from accurate pathological analyses, which provided a solid foundation for ultrasound-based video analysis of rectal cancer.

Image quality and interpretability: Image quality and interpretability are one of the main challenges when it comes to diagnosing rectal cancer ultrasound by endoscopic ultrasound means with deep learning analysis. This challenge is particularly highlighted in the analysis of video data, which is more complex and difficult to process and understand than static ultrasound images. The endoscopic ultrasound video data for rectal cancer presented in this paper contains time-varying image sequences that provide a dynamic view of cancer tissues and structures under ultrasound. This dynamic information is crucial for understanding the nature of the cancer or other physiological processes (e.g., blood flow dynamics at proliferations, etc.). However, video data introduces additional complexity because changes in the time dimension must be taken into account. In contrast, still image analysis only has to deal with data at a single point in time, making the task technically simpler. Therefore, the video form of RCEUV-207 requires more specific processing tools than simple still image recognition as in the past.

Time correlation for non-static analyses: Video data undoubtedly adds additional information in the time dimension compared to ultrasound image data with accurate labelling. How to effectively utilise this temporal information and extract this feature information useful for rectal cancer risk assessment constitutes a challenge in terms of ultrasound video-based analysis. Ultrasound is more dependent on the subjective experience of the operator, and therefore, it does not exhibit complete regularity in the form of video data. In RCEUV-207, even though we tried as much as possible to keep the lesion area in view for as long as possible, mismatches were inevitable due to human jitter. Therefore, how to build a dynamic analysis model for RCEUV-207 from the time dimension has also become one of the important challenges in this field.

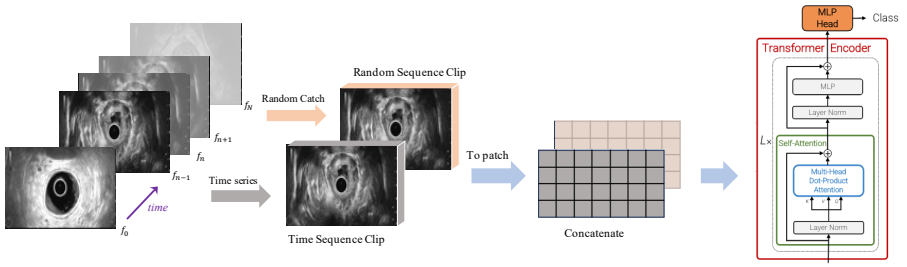

Fig. 2. The overall framework of the proposed method in this paper. The input video frames will be encoded and integrated from the perspectives of time series and random sequences, respectively. The concatenated image patches are sequentially fed into the Transformer Encoder network[21].

3 Method

To address the above challenges, this paper proposes a novel rectal cancer video assessment network (RCVA-Net), as shown in Figure 2. It effectively mitigates the uncertainty of ultrasound video-based analysis for rectal cancer brought about by the first and second challenges. Moreover, we propose a dynamic time-series fusion video feature extraction network to address the third challenge. Specifically, our motivation comes from how to identify data relationships in the temporal dimension of ultrasound videos and capture the link between previous and subsequent frames. Based on this be able to give a global view of the video data to better mitigate the problems mentioned in the third challenge. For this purpose, we have introduced two data encoding methods as follows respectively.

3.1 Neighbor Sequence Encoding Module(NSEM)

The proposed Neighbor Sequence Encoding Module is shown in Figure 3. Assuming that a video contains N frames, we use 3 consecutive frames as the basic data unit. Since ultrasound images are single-channel grey-scale maps, data units with temporal connections can be directly spliced together to obtain a $3\times224\times224$ tensor. Immediately after that, we use an image patching means to collate the $3\times224\times224$ data units into a 196×768 2-D matrix[3]. Due to the connectivity between the front and back frames, for a video containing N frames, we can obtain N 2-D matrices. For the head and tail of videos, we use them as the front and back frame neighbourhoods respectively to be able to be consolidated into data units. The re-encoded N 2-D matrices can reflect at a high level the connections between the front and back frames of the video and provide the basis for data preparation.

3.2 Random Sequence Encoding Module(RSEM)

In order to dynamically consider the global video data and to provide a comprehensive view of the data for designing the model, we propose a coding module for

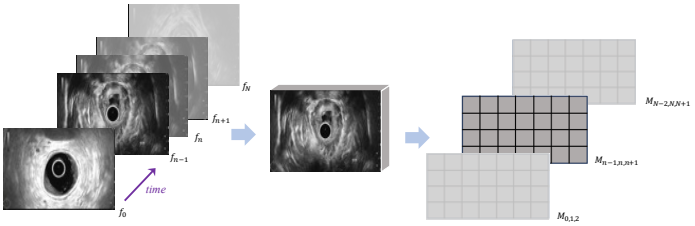

Fig. 3. The overall framework of the proposed method in this paper. The input video frames will be encoded and integrated from the perspectives of time series(**NSEM**) and random sequences(**RSEM**). The concatenated image patches are sequentially fed into the Transformer Encoder network.

stochastic grasping. It is composed of three frames of video footage grabbed, but unlike NSEM, the three frames constituting RSEM are independently sampled randomly along the time axis. This can be mathematically described as follows: for each frame f_j (where $j = 1, 2, 3$), a random index i_j is selected from the set $\{1, 2, \ldots, N\}$ without replacement. This ensures that each frame is uniquely and randomly selected, contributing to a comprehensive and diverse data representation. This suggests that RSEM will stochastically bring a global view to the feature learning process. For video data containing N frames, the randomised grab will result in taking $\lfloor N/3 \rfloor$ as the whole random data unit, where $\lfloor \cdot \rfloor$ denotes rounding down to the nearest whole number. For each of the three data frames obtained from the grabbing, we use the data unit integration method in section 3.1 for processing.

3.3 Feature Encoder Network

For the N 2D matrices obtained in the time series dimension with the $\lfloor N/3 \rfloor$ 2D matrices obtained from random grabs, we re-spliced them and fed them into the feature coding network. In this paper, we use Transformer Encoder[20] for feature computation. In RCAV-Net, we use ViT-b as the backbone for the feature extraction. Specifically, an RSEM output is concatenated after every three NSEM output tensors. The alternating use of NSEM and RSEM data loading methods enables the feature encoding network to acquire information on the temporal dimension by having a global view. In the input stage, we add the unique heat vector containing category information for model training. Instead of using the random sequence coding module, the temporal order data coding will be directly utilised for model inference.

Implementation Details: We use the ViT-b implementation pre-trained on ImageNet-22k to initialise the feature encoding network and act as the backbone network. The RCEUV-207 is divided into a training set and a test set at a ratio of 5:1 and conducted ten-fold cross-validation. The cross-entropy loss is used to control the model to learn the data categories. All videos used for training were

subjected to basic data augmentation strategies such as randomness flipping, cropping, and resizing. Our method was trained on torch 1.18, using Adam as an optimiser and setting weight decay to 0.0001, with 200 iterations at a setting of the batch of 2. All model training was performed on 2 Tesla V100(32G).

4 Experiments and Results

Evaluation Metrics: For the target task of this article, we use scenario-based classification therapy to evaluate the performance of the baseline model. Including accuracy, precision, recall and specificity.

4.1 Benchmarking

We benchmarked the RCEUV-207 from the perspective of 2D static image classification and video-level classification, respectively. We compared classical image classification methods like ResNet-50[6], EfficientNet[17] ,ViT[3] ,Swin[12] etc. and also used video classification methods on top of these algorithms like I3D[2] ,SlowOnly[4],TSM[8] , and Video-Swin[13] . The benchmarking results are shown in Table 1. In terms of accuracy, image-based classification methods are higher than video classification models, this is because static image analysis deals with each image frame individually without considering the complex relationships between time series. Comparatively, video classification models try to capture the temporal dynamics in the video, which increases the complexity of the model and may lead to lower accuracy in some cases, which is more relevant to real-world scenarios. As for video-level classification, it is easy to find that even for video-swin, which has strong feature extraction capabilities for video understanding, performing video-level classification of endoscopic ultrasound video for rectal cancer is still challenging, which only achieves 66.7% of top-1. In contrast, the RCVA-Net proposed in this paper is able to significantly improve the baseline results, achieving 77.2% of video-level accuracy. This provides a new baseline score for endoscopic ultrasound video-based rectal cancer classification.

4.2 Ablation Study

In this paper, two data encoding mechanisms, NSEM and RSEM, are proposed to ensure that the global dynamic representation of the video data is maintained while accurately capturing key features in continuous time variation. The results of the ablation experiments in Table 2 demonstrate that the performance of the model obtains significant improvement when both NSEM and RSEM are utilised. Specifically, NSEM provides the model with a correlation of video objects in a near-continuous process in the video, while RSEM complements the in-depth understanding of the relationships between the global scenes of the video and the changes in the temporal dimension. This two-pronged strategy allows the model to not only grasp the information of each frame but more importantly, to understand how these frames evolve and interrelate over time.

Table 1. Benchmark Results on RCEUV-207 Dataset

Model	Type	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)
ResNet-18[6]	Image	77.1	84.3	87.1	44.5
ResNet-50	Image	78.1	64.7	78.5	76.9
EfficientNet[17]	Image	78.4	93.3	63.6	94.4
ViT-b[3]	Image	81.9	85.7	70.5	91.3
Swin-t[12]	Image	80.0	76.9	66.7	88.0
SlowOnly(RN50)[4]	Video	38.9	37.5	47.3	28.5
TSM[8]	Video	39.0	29.6	57.1	26.9
SlowFast(RN101)[4]	Video	33.4	32.2	66.7	16.0
R(2+1)D[18]	Video	33.4	42.8	37.5	25.0
I3D(RN50)[2]	Video	27.8	38.9	28.0	26.6
Video-Swin-t[13]	Video	66.7	69.2	47.3	80.9
RCVA-Net(Ours)	Video	77.2	80.8	84.0	66.7

In addition, the ablation experiments demonstrated the unique roles and complementary nature of NSEM and RSEM in the model, with the absence of either element resulting in a degradation of the model's performance. This data loading and processing method not only enhances the model's ability to understand video data but also provides new ideas and methods for future video analysis and processing techniques.

Table 2. Ablation experimental results

	NSEM	RSEM	Accuracy (%)
Baseline	✓		66.05
		✓	51.23
RCAV-Net	✓	✓	77.21

5 Conclusion

In this paper, we first successfully proposed and labelled the first dataset for endoscopic ultrasound analysis of rectal cancer, which covers 207 cases and 5 different categories. The creation of this dataset provides a valuable resource for an in-depth understanding of endoscopic ultrasound image characterisation and classification of rectal cancer. Given the unique challenges of endoscopic ultrasound data analysis for rectal cancer, including the diversity of image qualities, the complexity of time-series data, and the problem of recognising subtle differences between categories, we propose a novel deep learning framework, RCAV-Net. RCAV-Net is designed to enhance the video classification model's ability to recognise features of rectal cancer by fusing continuous-frame and random-frame features. which in turn improves the accuracy of disease detection. Through a

series of benchmark tests on the proposed dataset, RCAV-Net demonstrates its significant advantages over existing state-of-the-art methods for the task of endoscopic ultrasound classification of rectal cancer.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant 82261138629 and 12326610; Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515010688; Shenzhen Municipal Science and Technology Innovation Council under Grant JCYJ20220531101412030

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ang, T.L., Kwek, A.B.E., Wang, L.M.: Diagnostic endoscopic ultrasound: technique, current status and future directions. Gut and Liver **12**(5), 483 (2018)
2. Carreira, J., Zisserman, A.: Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. pp. 6299–6308 (2017), https://openaccess.thecvf.com/content_cvpr_2017/html/Carreira_Quo_Vadis_Action_CVPR_2017_paper.html
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., others: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
4. Feichtenhofer, C., Fan, H., Malik, J., He, K.: SlowFast Networks for Video Recognition. pp. 6202–6211 (2019), https://openaccess.thecvf.com/content_ICCV_2019/html/Feichtenhofer_SlowFast_Networks_for_Video_Recognition_ICCV_2019_paper.html
5. Glynne-Jones, R., Wyrwicz, L., Tiret, E., Brown, G., Rödel, C.d., Cervantes, A., Arnold, D.: Rectal cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Annals of Oncology **28**, iv22–iv40 (2017), publisher: Elsevier
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. pp. 770–778 (2016), https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html
7. Horvat, N., Carlos Tavares Rocha, C., Clemente Oliveira, B., Petkovska, I., Gollub, M.J.: MRI of rectal cancer: tumor staging, imaging techniques, and management. Radiographics **39**(2), 367–387 (2019), publisher: Radiological Society of North America
8. Lin, J., Gan, C., Han, S.: TSM: Temporal Shift Module for Efficient Video Understanding. pp. 7083–7093 (2019), https://openaccess.thecvf.com/content_ICCV_2019/html/Lin_TSM_Temporal_Shift_Module_for_Efficient_Video_Understanding_ICCV_2019_paper.html
9. Lin, Y., Kou, S., Nie, H., Luo, H., Eltahir, A., Chapman, W., Hunt, S., Mutch, M., Zhu, Q.: Deep learning based on co-registered ultrasound and photoacoustic imaging improves the assessment of rectal cancer treatment response. Biomed. Opt. Express **14**(5), 2015–2027 (May 2023). https://doi.org/10.1364/BOE.487647, https://opg.optica.org/boe/abstract.cfm?URI=boe-14-5-2015
10. Liu, P., Zhang, J., Wu, X., Liu, S., Wang, Y., Feng, L., Diao, Y., Liu, Z., Lyu, G., Chen, Y.: Benchmarking supervised and self-supervised learning methods in a large ultrasound muti-task images dataset. IEEE Journal of Biomedical and Health Informatics pp. 1–12 (2024). https://doi.org/10.1109/JBHI.2024.3382604

11. Liu, S., Liu, Y., Xu, X., Chen, R., Liang, D., Jin, Q., Liu, H., Chen, G., Zhu, Y.: Accelerated cardiac diffusion tensor imaging using deep neural network. Physics in Medicine & Biology **68**(2), 025008 (2023)
12. 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. pp. 10012–10022 (2021)
13. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3202–3211 (2022)
14. Naito, Y., Tsuneki, M., Fukushima, N., Koga, Y., Higashi, M., Notohara, K., Aishima, S., Ohike, N., Tajiri, T., Yamaguchi, H., et al.: A deep learning model to detect pancreatic ductal adenocarcinoma on endoscopic ultrasound-guided fine-needle biopsy. Scientific reports **11**(1), 8454 (2021)
15. Santiago, I., Figueiredo, N., Parés, O., Matos, C.: Mri of rectal cancer—relevant anatomy and staging key points. Insights into Imaging **11**, 1–21 (2020)
16. Siddiqui, A.A., Fayiga, Y., Huerta, S.: The role of endoscopic ultrasound in the evaluation of rectal cancer. In: International Seminars in Surgical Oncology. vol. 3, pp. 1–7. Springer (2006)
17. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
18. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A Closer Look at Spatiotemporal Convolutions for Action Recognition. pp. 6450–6459 (2018), https://openaccess.thecvf.com/content_cvpr_2018/html/Tran_A_Closer_Look_CVPR_2018_paper.html
19. Uberoi, A.S., Bhutani, M.S.: Has the role of eus in rectal cancer staging changed in the last decade? Endoscopic ultrasound **7**(6), 366–370 (2018)
20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: 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), https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
22. Wiersema, M.J., Harewood, G.C.: Endoscopic ultrasound for rectal cancer. Gastroenterology Clinics **31**(4), 1093–1105 (2002)
23. Zhang, J., Chen, Y., Zeng, P., Liu, Y., Diao, Y., Liu, P.: Ultra-attention: automatic recognition of liver ultrasound standard sections based on visual attention perception structures. Ultrasound in Medicine & Biology **49**(4), 1007–1017 (2023)
24. Zhao, G., Kong, D., Xu, X., Hu, S., Li, Z., Tian, J.: Deep learning-based classification of breast lesions using dynamic ultrasound video. European Journal of Radiology **165**, 110885 (2023). https://doi.org/https://doi.org/10.1016/j.ejrad.2023.110885, https://www.sciencedirect.com/science/article/pii/S0720048X23001997