Harnessing Temporal Information for Precise Frame-Level Predictions in Endoscopy Videos

Pooya Mobadersany[1](✉), Chaitanya Parmar[1], Pablo F. Damasceno[1], Shreyas Fadnavis[1], Krishna Chaitanya[1], Shilong Li[1], Evan Schwab[2], Jaclyn Xiao[3], Lindsey Surace[1], Tommaso Mansi[1], Gabriela Oana Cula[1], Louis R. Ghanem[1], and Kristopher Standish[1](✉)

[1] Janssen R&D, LLC, a Johnson & Johnson Company
{pmobader, kstandis}@its.jnj.com
[2] Epic Sciences, San Diego, CA, USA
[3] University of California, San Francisco, CA, USA

Abstract. Camera localization in endoscopy videos plays a fundamental role in enabling precise diagnosis and effective treatment planning for patients with Inflammatory Bowel Disease (IBD). Precise frame-level classification, however, depends on long-range temporal dynamics, ranging from hundreds to tens of thousands of frames per video, challenging current neural network approaches. To address this, we propose EndoFormer, a frame-level classification model that leverages long-range temporal information for anatomic segment classification in gastrointestinal endoscopy videos. EndoFormer combines a Foundation Model block, judicious video-level augmentations, and a Transformer classifier for frame-level classification while maintaining a small memory footprint. Experiments on 4160 endoscopy videos from four clinical trials and over 61 million frames demonstrate that EndoFormer has an AUC=0.929, significantly improving state-of-the-art models for anatomic segment classification. These results highlight the potential for adopting EndoFormer in endoscopy video analysis applications that require long-range temporal dynamics for precise frame-level predictions.

Keywords: Camera localization · Video analysis · Foundation Model · Transformer · Temporal model · Inflammatory bowel disease · Endoscopy

1 Introduction

Crohn's Disease (CD) is a chronic immune-mediated disease characterized by transmural inflammation and mucosal ulceration throughout the intestinal tract [1]. In clinical trials, endoscopy videos are acquired to score CD severity, typically performed using the Simple Endoscopic Score for Crohn's Disease (SES-CD) [14] over five anatomic segments: rectum (RM), left colon/sigmoid (LC), transverse colon (TC), right colon (RC), ileum (IL). Automated video segmentation to anatomic segment classes is pivotal for automating SES-CD, possibly reducing inter-rater variability, and enhancing the scoring system [20].

Several methods for automating video segmentation rely on first identifying the camera location relative to some point (e.g. the beginning of the video) using powerful methods such as Simultaneous Localization and Mapping (SLAM) [7] or Optical Flow (OF) [8,16]. After the camera location is fully mapped, segments are identified by applying a fixed template to split the overall distance travelled by the camera into regions [17,28]. In practice, these template-based approaches can be unreliable because they do not account for inter-patient variations in segment length and bowel elasticity.

To avoid these issues, some methods have focused on the direct classification of frames into the anatomic segments, mostly using Convolutional Neural Networks (CNN) [3, 10, 22] or Long Short-Term Memory (LSTM) [11, 12, 19, 26] networks. The typically large length of these recordings, which can exceed 60,000 frames (∼30 mins at 30 frames per second (fps)), has so far hampered the use of methods able to capture long-range temporal dynamics, which we hypothesize to be the key for the identification of landmarks needed for accurate segmentation.

Recently, the application of Transformers [25] in various domains has showcased their exceptional ability to capture long-range dependencies [2, 9, 27] through self-attention mechanisms. Unlike CNNs and LSTMs that operate with fixed window sizes, Transformers simultaneously consider all positions in the input sequence, enabling the modeling of long-range dependencies. This strength comes at a significant memory cost, however, making them impractical for long endoscopy videos.

Motivated by these challenges, we present EndoFormer, aiming to strike a balance between computational feasibility and high frame-level performance via long-range consideration. To do so, we leverage a Foundation Model as a frame-level encoder, utilizing state-of-the-art self-supervised learning to learn high-quality frame-level representations. Furthermore, we employ a Transformer to capture inter-frame spatiotemporal relations across the entire video for the downstream task of anatomic segment classification. The main contributions of this paper can be summarized as follows:

- We propose a Transformer architecture for scalable segment prediction that incorporates long-range spatiotemporal information from endoscopy videos;
- We train a Foundation Model using over 61 million endoscopy frames, the largest dataset of its kind reported to date;
- We propose a set of video-level augmentations to improve model's robustness and generalization by mimicking real-world scenarios;
- We evaluate our model using two clinical trial datasets and improve robustness and generalization for automatic frame classification into SES-CD defined regions.

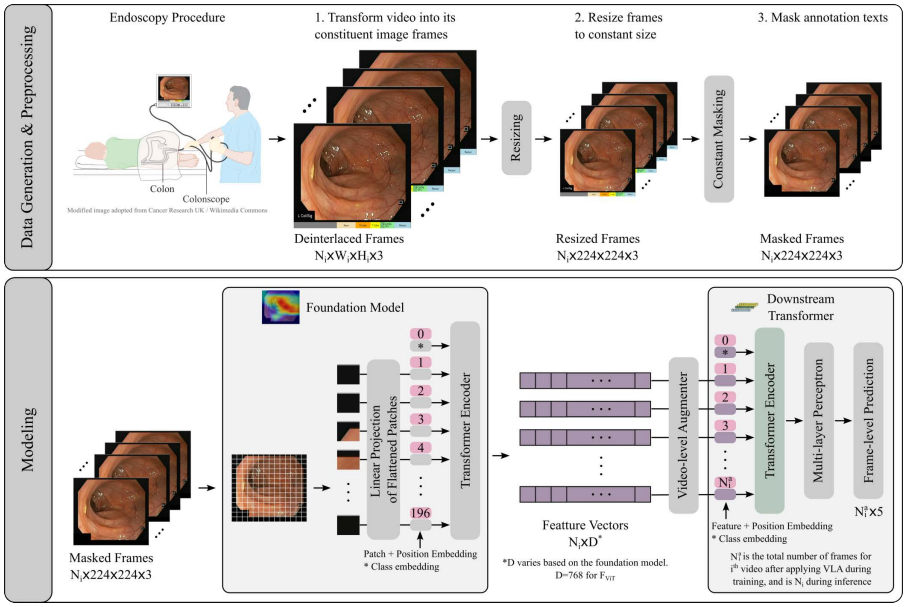

Fig. 1. EndoFormer pipeline. Data Generation & Preprocessing (top): each video is converted into deinterlaced image frames at 30 fps; each video might have different number of frames (N_i), width (W_i), height (H_i), so they are resized to $224 \times 224 \times 3$; constant masking is applied to all frames to prevent data leakage from text annotations already available in video frames. Modeling (bottom): masked frames are fed into the Foundation Model that encodes them into dense, feature-rich representations (embeddings). Video-level Augmenter applies random splits and/or reversals during EndoFormer model training to make the model robust against the real-world conditions of the endoscopy video recordings. Downstream Transformer leverages self-attention to effectively handle long-range dependencies while learning to classify each frame to one out of 5 anatomic segments.

2 Methodology

2.1 Framework Overview

Our proposed framework, EndoFormer, consists of three components (Fig. 1): a Foundation Model (FM) to extract enriched spatial features from frames; a Video-level Augmenter (VLA) to enhance the model's adaptability to various real-world conditions, and a Downstream Transformer which, by leveraging the encoded features, analyzes the video by considering both local and global temporal relations between frames to perform the downstream classification task without the large memory footprint of other temporal models.

Foundation Model: The FM encodes input data into a dense feature-rich spatial representation. Building on its successful benchmarking in various tasks, we employed DINOv2 [18] for pre-training the encoder through self-supervised

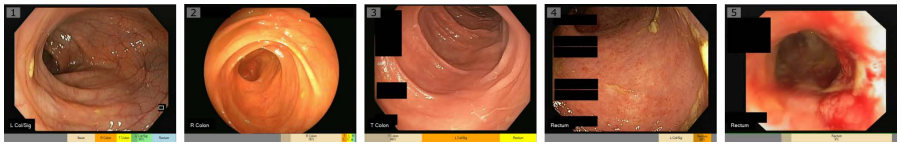

Fig. 2. Example frames of CD endoscopy videos. (1) example showing all 5 segments. In this data, the segment label as well as a progress bar indicating the location of the endoscope are available as text at the bottom of the image. The gray area at the beginning of the bottom progress bar represents the unannotated frames during forward navigation from the anus to the terminal ileum. The ground truth segment annotations for the SES-CD are made during the withdrawal phase of an endoscopy. (2-5) examples of recordings missing some of the segments.

learning. DINOv2 employs a student-teacher model paradigm, where both models share the same architecture and utilize knowledge distillation. The student model is trained on noisy variants of global views ($224 \times 224 \times 3$), while the teacher model is exposed to variants of local views ($96 \times 96 \times 3$). The training objective involves DINO [4], iBOT [29], and KoLeo [21] losses, which are collectively trained using a student-teacher setup.

$$\mathcal{L}_{DINOv2} = \mathcal{L}_{DINO} + \mathcal{L}_{iBOT} + \mathcal{L}_{KoLeo} \tag{1}$$

The outputs of the teacher model undergo centering via batch mean, and the weights are updated using an exponential moving average of the student model weights. Furthermore, we applied additional regularization techniques recommended in [18] during the training process.

Video-level Augmenter: VLA was designed to mimic real-world conditions that cause videos to start or end in non-standard colon segments. During the endoscopy, physicians insert an endoscope into the anus and advance it to the terminal ileum (in CD), cecum (in Ulcerative Colitis (UC), a different subtype of IBD), or a location determined based on factors such as disease severity or protocol (colonoscopy or flexible sigmoidoscopy). At any point during this insertion, the video recording may be turned on and after the region of interest is located, the clinician is instructed to withdraw the endoscope. These reasons lead to a wide variety of "ground truth" segments being present in clinical datasets (Fig. 2). To account for this variability and to avoid biases during training we devised a simple step whereby videos are randomly split and/or reversed prior to being loaded by the downstream Transformer (see Algorithm 1). The constraints $\pm\frac{L}{2}$ applied in lines 3 and 4 of Algorithm 1 were to ensure the augmented subset F_i^{aug} of F_i will have at least L rows after applying the random split. A lower value of L may result in a small number of selected frames. In our experiments, we set L to 2 for a more aggressive form of augmentation, while capturing a wider range of video lengths during model training. The VLA has a negligible effect on the training duration as it applies augmentations to the FM-extracted features that map to the video frames.

Algorithm 1: Video-level Augmenter

Input: Matrix of embeddings in i^{th} video during training (F_i)
Output: Augmented subset of F_i (F_i^{aug})

$r_i^{split} \leftarrow$ Random float number between 0 and 1;
if $r_i^{split} \geq 0.5$ **then**
  $r_{i,start}^{split} \leftarrow$ Random integer between 0 and $\frac{1}{2}N_i - \frac{L}{2}$ where $L = \{l \in \mathbb{Z} \mid l > 0 \; and \; l \leq N_i\}$
  $r_{i,end}^{split} \leftarrow$ Random integer between $\frac{1}{2}N_i + \frac{L}{2}$ and N_i
  $F_i^{aug} \leftarrow$ Rows $r_{i,start}^{split}$ to $r_{i,end}^{split}$ from F_i
else
  $F_i^{aug} \leftarrow F_i$
end
$r_i^{reverse} \leftarrow$ Random float number between 0 and 1;
if $r_i^{reverse} \geq 0.5$ **then**
  $F_i^{aug} \leftarrow$ Reverse the order of embeddings (rows) in F_i^{aug}
end
return F_i^{aug}

Downstream Transformer: The frame-level prediction problem is mathematically represented as $\widehat{C}_{i,j} = S(f_{i,j})$, where S is the classification function that maps the input embedding $f_{i,j}$ of the j^{th} frame from the i^{th} video ($\mathcal{V}_i$) to its corresponding anatomic segment class $\widehat{C}_{i,j}$. To accurately predict anatomic segments, the classification function S must consider neighboring embeddings and capture both short- and long-range dependencies. This is crucial because anatomic segments demonstrate a biological order, and accurately incorporating information from neighboring embeddings helps improve prediction accuracy.

To leverage long-range temporal relations for frame classification, we employed a Transformer architecture. The input matrix per video consisted of N_i^a frames with D-dimensional feature representing each frame, where N_i^a is the number of frames for $\mathcal{V}_i$ after applying VLA during training, and is the total number of frames (N_i) for $\mathcal{V}_i$ available during inference. This matrix was passed through the Transformer, followed by a multi-layer perceptron (MLP) for anatomic segment classification of each frame. The downstream Transformer was trained using cross-entropy loss.

2.2 Datasets and Preprocessing

We used data from four clinical trials, involving 1847 patients and 4160 videos, for training, validation, and testing. Two datasets comprised patients with Crohn's disease (CD) (ClinicalTrials.gov IDs: NCT03464136, NCT02877134), while the other two datasets included patients with ulcerative colitis (UC) (ClinicalTrials.gov IDs: NCT02407236, NCT01959282). UC datasets lack anatomic segment labels due to the global disease severity scoring used for UC, which

does not include the SES-CD criteria. Thus, UC datasets were only used for FM pre-training, not for the downstream task. We obtained high-quality ground truth for our downstream transformer model through a hybrid approach. This involved automatic text extraction using an OCR-based algorithm [24], followed by meticulous manual review and refinement. This process was applied to all frames from the CD clinical trial endoscopy videos. Approximately 50% of these frames did not have textual annotations for anatomic segments, mainly due to their placement in the forward path. These frames were categorized as "Unknown" and excluded from the downstream task. The remaining frames were mapped to standardized anatomic segment labels: IL, RC, TC, LC, and RM. We allocated over 21 million labeled frames from 1335 videos of 753 patients for Downstream Transformer training, validation, and testing, using a 70:10:20% split ratio, respectively. The 20% CD data allocated for Downstream Transformer testing was excluded from FM training.

2.3 Settings and Metrics

EndoFormer: Our proposed EndoFormer pipeline consists of a Foundation Model, Video-level Augmenter, and a Downstream Transformer. For the FM (F_{ViT}) we used ViT-B/16 [6] and pre-trained it by utilizing the DINOv2 [18] model training paradigm. We used a batch size of 256, cosine decayed learning rate of 8×10^{-4} and Adam optimizer [13] for 15 epochs on 4 NVIDIA A10G GPUs (total training = 8 days). The learning rate was warmed up for the first 10% of the iterations and then proportionally decayed to zero until the last iteration. We used this encoder to extract features ($D = 768$) for each frame. The Downstream Transformer has 4 layers and 8 self-attention heads in each layer and a dropout of 0.25 for the Transformer layer and 0.5 for the final classification layer. Adam optimizer with learning rate 10^{-5} along with a weight decay of 10^{-6} was used for training the EndoFormer model, and the model with highest AUC on validation set was selected for model evaluation on test set.
Comparison with State-of-the-Art: We compare EndoFormer to best-performing algorithms from four categories: template-based, CNN-based, LSTM-based and Foundation-based models. For comparability, we have optimized all these models on the same training dataset used for EndoFormer downstream training. For template-based models [28], we used Gunnar FarneBack's dense OF method [8] to estimate the cumulative distance traveled by camera after which a pre-defined template was used to map frame location to anatomic segments. For CNN-based models [3], we trained a ResNet101 using a fully-supervised learning approach and a weighted cross entropy loss to classify frames into anatomic segments. For LSTM-based models [12], we replace our downstream Transformer classifier with an LSTM layer while using the encoded features from the F_{ViT} model. For Foundation-based models [6], we keep our encoded features that were generated by the ViT-based Foundation Model (F_{ViT}) and replace the downstream Transformer classifier with a linear layer.
Evaluation Metrics: We used AUC, F1, Accuracy and Adjacent accuracy (Adj. accuracy) as evaluation metrics for the downstream task. We investigated both

Model	mAvg AUC (%)	mAvg F1 (%)	Accuracy (%)	Adj accuracy (%)
Endomapper [3]	67.5 ± 0.1	12.3 ± 0.0	29.4 ± 0.1	62.8 ± 0.1
ViT [6]	78.5 ± 0.0	45.7 ± 0.1	49.3 ± 0.1	77.4 ± 0.1
Template-based [28]	69.8 ± 0.0	50.6 ± 0.1	53.6 ± 0.1	93.5 ± 0.0
TMRNet [12]	85.7 ± 0.0	58.4 ± 0.1	61.6 ± 0.1	87.5 ± 0.0
EndoFormer *(ours)*	**92.9 ± 0.0**	**70.5 ± 0.1**	**72.8 ± 0.1**	**95.4 ± 0.0**

Table 1. Frame-level performance of different models on CD test set.

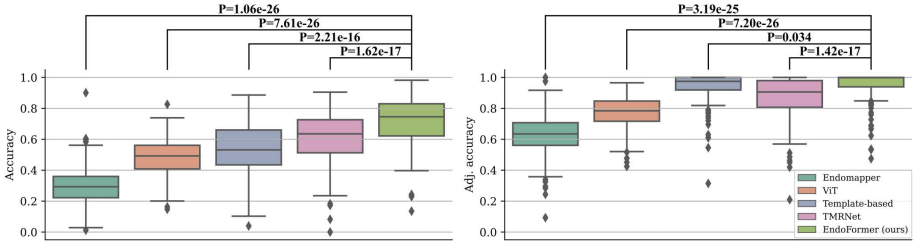

Fig. 3. Comparison between patient-level Accuracy (left) and Adjacent accuracy (right) for each model on CD test set.

frame-level and patient-level performances. In the frame-level analysis we investigate the model performance across all frames in the test set, where we do a bootstrap of test set size with 100 iterations to obtain the mean and standard deviation (std) for AUC, F1, Accuracy, and Adj. accuracy. For patient-level analysis, overall Accuracy and Adj. accuracy are evaluated across all the frames for each patient. Wilcoxon signed-rank test was performed to compare the performance of different models; we chose this paired test because each method was evaluated using identical train/test sets.

3 Results and Discussion

Table 1 shows that EndoFormer outperforms previous methods for frame-level classification across all metrics. We find that fully supervised method and template-based have the lowest AUC performance, followed by foundation encoder methods using linear and LSTM downstream classifiers. Additionally, Fig. 3 shows EndoFormer's superior patient-level performance compared to other models (Wilcoxon signed-rank $P < 0.05$).

To better understand which component of EndoFormer was responsible for this performance gain, we performed a series of ablation studies. Table 2 illustrates a major decline in the EndoFormer performance upon substituting the F_{ViT} with a ResNet101 pre-trained on ImageNet whereas only a slight performance reduction was observed upon replacing F_{ViT} with CNN-based Foundation Model (F_{CNN}) with a ResNet101 backbone that was pre-trained using SimCLR [5] model training paradigm. The difference between the patient-level

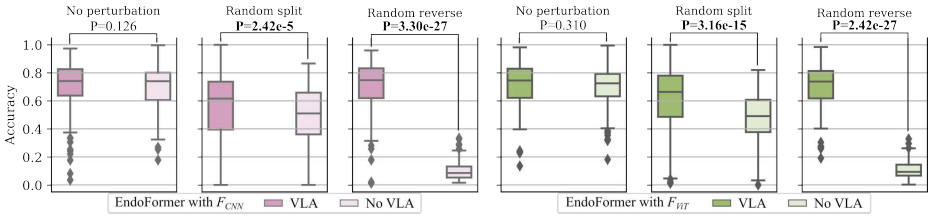

Fig. 4. Ablation study assessing the impact of removing Video-level Augmenter for EndoFormer with different type of FMs (left: F_{CNN}, right: F_{ViT}). Three inference conditions were tested: No perturbation (no random splits or reversals), Random split (random video-level splits), and Random reverse (random video-level reversals).

performance using each of these FMs was not statistically significant (Wilcoxon signed-rank $P = 0.829$ and 0.945 for Accuracy and Adj. accuracy, respectively); however, F_{ViT} is preferred here since its training time is ∼40% faster. Overall, these results illustrate the impact of the FM in EndoFormer performance; using F_{ViT} or F_{CNN} as the FM significantly improves the model performance compared to using an encoder pre-trained on ImageNet. Additionally, we examined the impact of VLA on the EndoFormer with F_{ViT} and F_{CNN} FMs. Fig. 4 demonstrates that the proposed video-level augmentations during training effectively mitigate the impact of random reversals and splits on patient-level performance during testing. When there is no perturbations during inference, Accuracies are similar with and without VLA (Wilcoxon signed-rank $P > 0.05$). However, under perturbations of random reversals or splits, the EndoFormer with VLA surpasses the EndoFormer without it (Wilcoxon signed-rank $P < 0.005$). In Fig. 5, the t-SNE plot displays embeddings from the final layer of the Downstream Transformer on the CD test set. The plot reveals well-clustered embeddings for each segment and closer clustering of neighboring segments. This demonstrates that the combination of the FM with a spatiotemporal Transformer in EndoFormer effectively captures both local and global representations.

4 Conclusions

We have developed a novel pipeline for frame-level classification of anatomic segments in endoscopy videos. We introduced a spatiotemporal model to classify

Backbone	(Pre)training mechanism	Accuracy (%)	Adj. accuracy (%)	(Pre)training Time
ResNet101	ImageNet	46.3 ± 20.3 (↓ 24.4)	78.0 ± 20.4 (↓ 17.1)	-
ResNet101	SimCLR	70.4 ± 17.7 (↓ 0.3)	94.9 ± 10.8 (↓ 0.2)	14 days
ViT	DINOv2	**70.7 ± 15.8**	**95.1 ± 9.6**	8 days

Table 2. Ablation study assessing the impact of substituting FM on patient-level performance in EndoFormer. Last row is our proposed Foundation Model (F_{ViT}).

Fig. 5. t-SNE analysis of the generated embeddings in the final Transformer layer of EndoFormer (before the classification layer) for CD test set.

segments and developed video-level augmentations to make the model robust to different real-world scenarios. Our results show that: 1) extracting features from a domain-specific FM improves performance over more general pre-trained models, 2) leveraging Transformers to incorporate local and global information over many frames improved model performance, and 3) applying the proposed video-level augmentations during training makes the model resilient against random perturbations introduced with different video splits and directions. Future work will investigate the generalization of anatomic segment classification to other types of IBD such as UC for novel endpoint development [15, 23]. Full automation will require automatically detecting the withdrawal path of the endoscope as it navigates through the colon. Additionally, we will investigate the impact of choosing different downsampling rates for the frames in the model's performance. Segmenting endoscopy videos is useful for automating the SES-CD and developing advanced novel endpoints for localized disease scoring, though this approach can be leveraged across a variety of tasks where global information can improve local, frame-level classification.

Disclosure of Interests. All authors were employees of Janssen R&D, LLC, when conducting this research, and may own company stock/stock options.

References

1. Ananthakrishnan, A.N., Kaplan, G.G., Ng, S.C.: Changing global epidemiology of inflammatory bowel diseases: sustaining health care delivery into the 21st century. Clinical Gastroenterology and Hepatology **18**(6), 1252–1260 (2020)
2. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6836–6846 (2021)
3. Azagra, P., Sostres, C., Ferrández, Á., Riazuelo, L., Tomasini, C., Barbed, O.L., Morlana, J., Recasens, D., Batlle, V.M., Gómez-Rodríguez, J.J., et al.: Endomapper dataset of complete calibrated endoscopy procedures. Scientific Data **10**(1), 671 (2023)

4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=YicbFdNTTy
7. Durrant-Whyte, H., Bailey, T.: Simultaneous localization and mapping: part i. IEEE robotics & automation magazine **13**(2), 99–110 (2006)
8. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13. pp. 363–370. Springer (2003)
9. Gao, X., Jin, Y., Long, Y., Dou, Q., Heng, P.A.: Trans-svnet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24. pp. 593–603. Springer (2021)
10. Houwen, B.B., Hartendorp, F., Giotis, I., Hazewinkel, Y., Fockens, P., Walstra, T.R., Dekker, E., study group, P.: Computer-aided classification of colorectal segments during colonoscopy: a deep learning approach based on images of a magnetic endoscopic positioning device. Scandinavian Journal of Gastroenterology **58**(6), 649–655 (2023)
11. Jin, Y., Li, H., Dou, Q., Chen, H., Qin, J., Fu, C.W., Heng, P.A.: Multi-task recurrent convolutional network with correlation loss for surgical video analysis. Medical image analysis **59**, 101572 (2020)
12. Jin, Y., Long, Y., Chen, C., Zhao, Z., Dou, Q., Heng, P.A.: Temporal memory relation network for workflow recognition from surgical video. IEEE Transactions on Medical Imaging **40**(7), 1911–1923 (2021)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Koutroumpakis, E., Katsanos, K.H.: Implementation of the simple endoscopic activity score in crohn's disease. Saudi journal of gastroenterology: official journal of the Saudi Gastroenterology Association **22**(3), 183 (2016)
15. Lobatón, T., Bessissow, T., De Hertogh, G., Lemmens, B., Maedler, C., Van Assche, G., Vermeire, S., Bisschops, R., Rutgeerts, P., Bitton, A., et al.: The modified mayo endoscopic score (mmes): a new index for the assessment of extension and severity of endoscopic activity in ulcerative colitis patients. Journal of Crohn's and Colitis **9**(10), 846–852 (2015)
16. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: IJCAI'81: 7th international joint conference on Artificial intelligence. vol. 2, pp. 674–679 (1981)
17. Morlana, J., Tardós, J.D., Montiel, J.: Colonmapper: topological mapping and localization for colonoscopy. arXiv preprint arXiv:2305.05546 (2023)
18. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)

19. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
20. Rutgeerts, P., Reinisch, W., Colombel, J.F., Sandborn, W.J., D'Haens, G., Petersson, J., Zhou, Q., Iezzi, A., Thakkar, R.B.: Agreement of site and central readings of ileocolonoscopic scores in crohn's disease: comparison using data from the extend trial. Gastrointestinal Endoscopy **83**(1), 188–197 (2016)
21. Sablayrolles, A., Douze, M., Schmid, C., Jégou, H.: Spreading vectors for similarity search. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=SkGuG2R5tm
22. Saito, H., Tanimoto, T., Ozawa, T., Ishihara, S., Fujishiro, M., Shichijo, S., Hirasawa, D., Matsuda, T., Endo, Y., Tada, T.: Automatic anatomical classification of colonoscopic images using deep convolutional neural networks. Gastroenterology report **9**(3), 226–233 (2021)
23. Schwab, E., Cula, G.O., Standish, K., Yip, S.S., Stojmirovic, A., Ghanem, L., Chehoud, C.: Automatic estimation of ulcerative colitis severity from endoscopy videos using ordinal multi-instance learning. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization **10**(4), 425–433 (2022)
24. Smith, R.: An overview of the tesseract ocr engine. In: Ninth international conference on document analysis and recognition (ICDAR 2007). vol. 2, pp. 629–633. IEEE (2007)
25. 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)
26. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. pp. 2048–2057. PMLR (2015)
27. Yamazaki, K., Vo, K., Truong, Q.S., Raj, B., Le, N.: Vltint: visual-linguistic transformer-in-transformer for coherent video paragraph captioning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 3081–3090 (2023)
28. Yao, H., Stidham, R.W., Gao, Z., Gryak, J., Najarian, K.: Motion-based camera localization system in colonoscopy videos. Medical Image Analysis **73**, 102180 (2021)
29. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: Image BERT pre-training with online tokenizer. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=ydopy-e6Dg