Endoscopic Scoring and Localization in Unconstrained Clinical Trial Videos

Supplementary

1. Dataset Visualization

Dataset	Videos	Frames
Colonoscopic [4]	210	36534
SUN&SUN-SEG [2,5]	1018	159400
LDPolypVideo [3]	237	40186
Kvasir-Capsule [8]	5704	875940
yper-Kvasir [1]	1000	158892
CholecTriplet [6]	580	90444
LIMUC [7]	\	11276
Our Dataset (small)	556	13,900
Our Dataset (large)	86423	518538

Table 1. Details of Dataset

We provide the details of all datasets here. The Colonoscopic dataset includes 210 videos with a total of $36,534$ frames. The SUN and SUN-SEG datasets have $1,018$ videos with $159,400$ frames. The LDPolypVideo dataset consists of 237 videos with $40,186$ frames. The Kvasir-Capsule dataset contains $5,704$ videos with $875,940$ frames. The Hyper-Kvasir dataset comprises $1,000$ videos with $158,892$ frames. The CholecTriplet dataset includes 580 videos with $90,444$ frames. The LIMUC dataset has $11,276$ frames. The Clario dataset consists of 556 videos with $13,900$ frames. Our dataset includes $86,423$ videos with a total of $518,538$ frames.

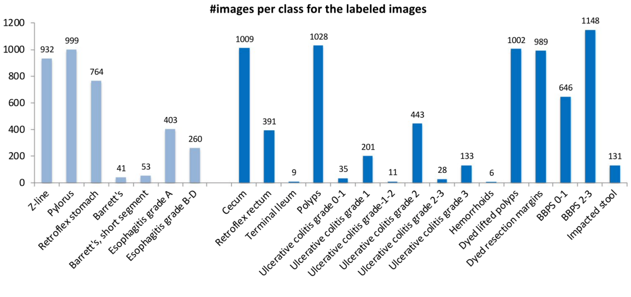


Figure 1. Dataset summary: HyperKvasir

In ulcerative colitis (Figure1), there are some confusing ones labeled as 0-1/1-2/2-3. Because in such cases, it is difficult to determine the exact class. Previous studies have shown important observer variation in assessing the degree of inflammation. In BBPS, four levels (0, 1, 2, 3) are grouped into two levels(0-1, 2-3). Therefore, we drop this dataset.

	Train + Validation	Test	Total
	85% from 479 patients	15% from 85 patients	564 patients
Mayo 0	5180	925	6105
Mayo 1	2588	464	3052
Mayo 2	1077	177	1254
Mayo 3	745	120	865
Total	9590	1686	11276

Table 2. LIMUC Dataset.

In Table 2 demosntate that LIMUC dataset is divided into training and validation sets, which consist of 85% of the data from 479 patients, and a test set, which consists of 15% of the data from 85 patients, making a total of 564 patients.

For Mayo 0, there are $5,180$ instances in the training and validation set, 925 in the test set, and $6,105$ in total. For Mayo 1, there are $2,588$ instances in the training and validation set, 464 in the test set, and $3,052$ in total. For Mayo 2, there are $1,077$ instances in the training and validation set, 177 in the test set, and $1,254$ in total. For Mayo 3, there are 745 instances in the training and validation set, 120 in the test set, and 865 in total. Overall, there are $9,590$ instances in the training and validation set, $1,686$ in the test set, and $11,276$ in total.

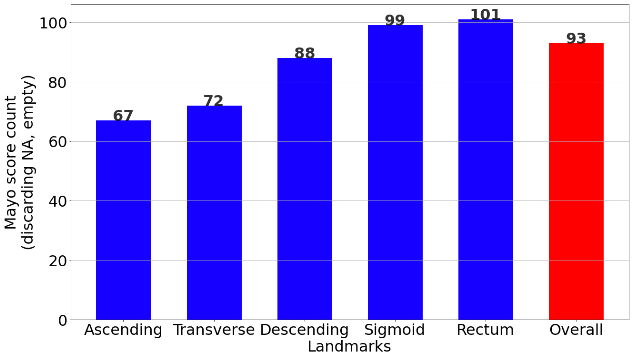


Figure 2. Our Dataset Distribution

The Figure 2 presents data from endoscopic experiments.

The ascending colon extends from the cecum to the hepatic flexure; the descending colon runs from the splenic flexure to the sigmoid colon. The rectum spans from the sigmoid colon to the anal canal. The sigmoid colon connects the descending colon to the rectum, while the transverse colon crosses horizontally between the hepatic and splenic flexures. The 556 video clips cover these segments, providing essential information for diagnosing and monitoring colorectal health.

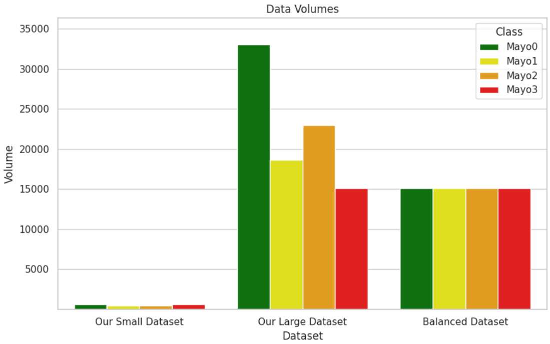


Figure 3. We compare three different settings in an imbalanced setting ablation study (Main paper, Table 4). The first version uses our small dataset, which is imbalanced and has less volume. The second version uses our larger dataset. The last version randomly drops the imbalanced part and merges a balanced version for Table 4.

2. Results Analysis

2.1. Slowfast Model:

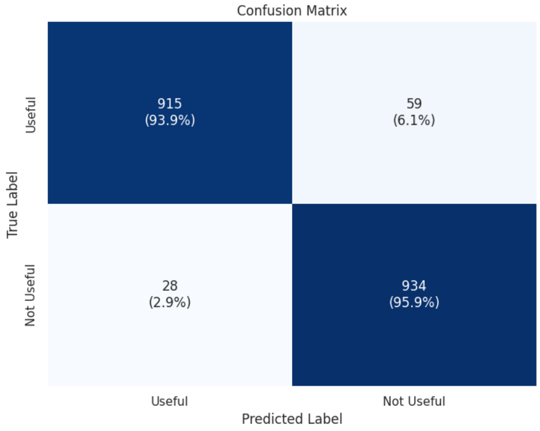


Figure 4. Confusion Matrix for Slow-fast Model in Active Learning

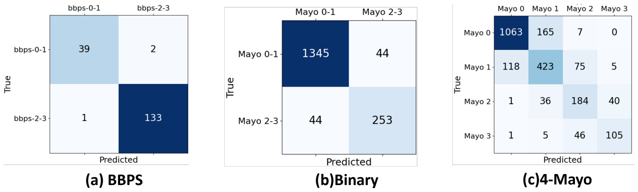


Figure 5. Confusion Matrix for HyperKvasir dataset: Binary; Binary Mayo scores; 4-class Mayo scores, from left to right.

We offer different confusion matrices with various methods. Details are provided here. Figure 4 is the Confusion Matrix for the SlowFast Model in Active Learning; Figure 5 is the Confusion Matrix. The confusion matrix reveals the following: True Positives 934, True Negatives: 915. In this classification task, the model performs well with a high number of correct predictions for both 'Useful' and 'Not Useful' labels.

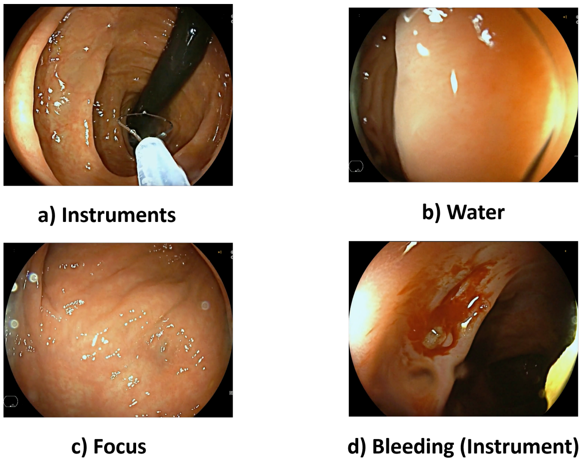
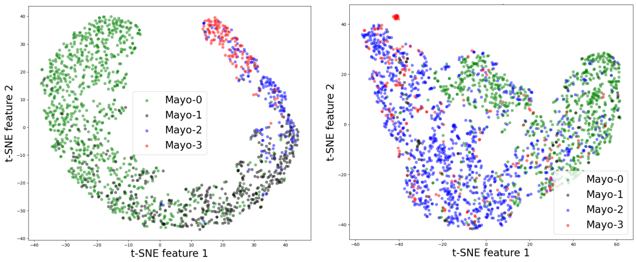


Figure 6. Failure cases



(a) LIMUC (b) Clario

Figure 7. t-SNE visualization of embeddings for the four-level Mayo score classes across two distinct datasets: (a) LIMUC and (b) Clario.

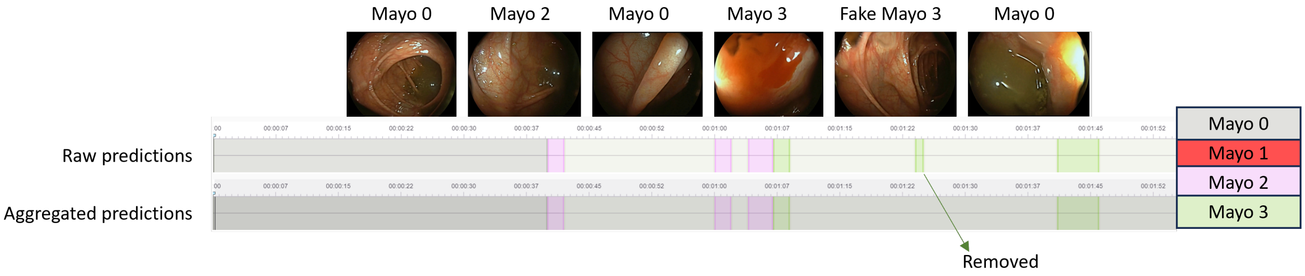


Figure 8. Label Studio User Interface Demo: Our user interacts with online Label Studio platform, where the interface is organized as follows: The first row displays the full-length video, the second row shows raw predictions, and the last row presents aggregated predictions, which help to filter out false predictions. (demo videos are available in the supplementary materials.)

2.2. Frame-based Model:

Beyond the ablation study in the main paper, Figure 6, a similar observation can be made from the t-SNE plots in Figure 7. This is likely due to the annotation being at the landmark level; even though a segment is annotated as Mayo-3, not the entire landmark consistently reflects Mayo-3 characteristics, potentially shifting between Mayo-2 and Mayo-3 due to varying inflammation points in the GI tract. This variability highlights the necessity for a video-level model that can account for these fluctuations, providing a more accurate representation of the disease state across different GI segments. In Table 3, we compare our approach with Multiple Instance Learning (MIL, a frame-based work). This approach is particularly useful in scenarios where only bag-level labels (sets of instances) are available. We integrate the MIL framework into our pipeline as a frame-based algorithm and then compare it with our video-based algorithm. Specifically, to integrate the MIL framework into an image classification algorithm, we implement an aggregation layer to combine instance-level feature representations into a single bag-level representation before making predictions. Experimental results demonstrate that MIL achieves similar performance to the X3D series model in binary cases (81.3%); however, when applied to a complex scenario, it drops significantly in accuracy(61% VS. 58%).

Task	Model	Test Accuracy (%)	AUC (weighted)	F1 (weighted)	Kappa (quadratic)
Binary	MIL	81.33	0.886	0.811	0.621
4-Class	MIL	58.37	0.725	0.327	0.447

Table 3. Comparisons with Multiple instance learning (MIL) on combined dataset.

2.3. Failure Cases

In Figure 6, we categorized all failure cases into the following groups: 1) Misclassifications, 2) Incorrect ground truth, 3) Instrument issues, and 4) Loss of focus, bubbles, and bleeding. We included visual examples for categories 2 to 4 to enhance understanding.

2.4. Demo

In Figure 8, we show our Label Studio user interface. Specifically, this demo is part of our end-to-end endoscopic scoring and localization system. The user only needs to pass raw data, and the model will automatically provide predictions and aggregated Mayo scores, which will be uploaded to Label Studio automatically. The user can check the final visualization and predictions. For example, in Label Studio, we merge all clips into a single video and then offer both individual predictions and aggregated predictions. Using the aggregation method, we drop the fake green (Mayo score 3). Finally, we use majority voting to obtain the video-level Mayo score prediction.

References

[1] Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, et al. Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific data*, 7(1):283, 2020. 1

[2] Ge-Peng Ji, Guobao Xiao, Yu-Cheng Chou, Deng-Ping Fan, Kai Zhao, Geng Chen, and Luc Van Gool. Video polyp segmentation: A deep learning perspective. *Machine Intelligence Research*, 19(6):531–549, 2022. 1

[3] Yiting Ma, Xuejin Chen, Kai Cheng, Yang Li, and Bin Sun. Ldpolypvideo benchmark: a large-scale colonoscopy video dataset of diverse polyps. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 387–396. Springer, 2021. 1

[4] Pablo Mesejo, Daniel Pizarro, Armand Abergel, Olivier Rouquette, Sylvain Beorchia, Laurent Poincloux, and Adrien Bartoli. Computer-aided classification of gastrointestinal lesions in regular colonoscopy. *IEEE transactions on medical imaging*, 35(9):2051–2063, 2016. 1

[5] Masashi Misawa, Shin-ei Kudo, Yuichi Mori, Kinichi Hotta, Kazuo Ohtsuka, Takahisa Matsuda, Shoichi Saito, Toyoki Kudo, Toshiyuki Baba, Fumio Ishida, et al. Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). *Gastrointestinal endoscopy*, 93(4):960–967, 2021. 1

[6] Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78:102433, 2022. 1

[7] Gorkem Polat, Haluk Tarik Kani, Ilkay Ergenc, Yesim Ozen Alahdab, Alptekin Temizel, and Ozlen Atug. Improving the computer-aided estimation of ulcerative colitis severity according to mayo endoscopic score by using regression-based deep learning. *Inflammatory Bowel Diseases*, 29(9):1431–1439, 2023. 1

[8] Pia H Smedsrud, Vajira Thambawita, Steven A Hicks, Henrik Gjestang, Oda Olsen Nedrejord, Espen Næss, Hanna Borgli, Debesh Jha, Tor Jan Derek Berstad, Sigrun L Eskeland, et al. Kvasir-capsule, a video capsule endoscopy dataset. *Scientific Data*, 8(1):142, 2021. 1