

Supplementary Materials: A Medical Data-Effective Learning Benchmark for Highly Efficient Pre-training of Foundation Models

Anonymous Authors

1 ADDITIONAL RESULTS

1.1 Pretraining Data Effectiveness Analysis

Table 1 provides a comprehensive analysis of the mIoU comparison for the ViT-Base model across various datasets and different proportions of pretraining data. The mIoU values are presented for pretraining data percentages of 5%, 10%, 20%, 33%, 50%, and 100%, with a comparison to fine-tuning on the initial parameters from ImageNet.

Table 1 indicates that the mIoU of pretraining datasets is generally higher than those fine-tuned using ImageNet parameters. This suggests that pretraining the model on our selected dataset effectively enhances overall model performance. Although for specific datasets (e.g., PolypGen, ImageCLEFmed), at certain pretraining data proportions (such as 5% and 20%), the mIoU may be slightly lower compared to the non-pretraining scenario, overall, training with pretraining data proves to be effective, highlighting the positive impact of pretraining on model performance improvement.

2 TEMPORAL PERFORMANCE ANALYSIS

Table 2, Table 3 and Table 4 provide in-depth insights into the temporal performance and training characteristics of the ViT-Base model in various contexts.

Table 2 provides an analysis of performance metrics during data processing at various similarity threshold values (η). It encompasses details on remaining data, ratios, and the specific time allocated to MedDEL (DEL) operations, illuminating the model’s behavior across a spectrum of η , which indicates that as the remaining data decreases, MedDEL (DEL) Time tends to increase.

In Table 3, the focus shifts to the temporal aspects of training, specifically analyzing the training time for the MAE based on the ViT-Base model. The results, encompassing different proportions of pretraining data and varying epochs, highlight a noteworthy observation: When diminishing the training data and concurrently amplifying the number of epochs proportionally, the training time demonstrates remarkable stability. Hence, utilizing a smaller dataset and conducting additional epochs in training is justifiable as a fair training strategy.

Table 4 delves into the temporal dynamics of the ViT-Base model’s training outcomes across diverse datasets and varying proportions of pretraining data. Noteworthy is the consistency in training times across different datasets at the same pretraining data proportion. This uniformity suggests a robust and steadfast performance of the model across a range of datasets and pretraining configurations.

3 MORE DISCUSSION

3.1 Analysis of data types

In our data-effective process, we categorize data into four types: disruptive data, invalid data, effective data, and core critical data.

Disruptive data, which includes irrelevant or erroneous information, hampers the performance of the model and needs to be eliminated to enhance training efficiency and the model’s generalization ability. Invalid data, offering no practical value to the task, must also be removed to prevent disruptions during training and inference, thereby improving efficiency. In contrast, effective data provides foundational material for model learning, aiding the model in better understanding the task and enhancing its performance. Finally, core critical data, which we are particularly interested in and wish to be a significant proportion, contains key features and information relevant to the task. Our goal is to preserve and fully utilize these data, ensuring the model excels in critical aspects of the task.

3.2 Further Details About the Dataset

Gastrointestinal Endoscopy Image and Video Datasets. The Gastrovision [9] dataset aims to improve the recognition of various gastrointestinal diseases, containing a total of 8,000 images. This dataset includes images of various anatomical landmarks, pathological abnormalities, and normal findings, covering lesions in the stomach, duodenum, esophagus, and other areas. Its diversity makes it a valuable resource for research on gastrointestinal diseases, suitable for both clinical auxiliary diagnosis and as a foundation for training deep learning and artificial intelligence algorithms.

Hyper-Kvasir [6] is one of the largest publicly available gastrointestinal image datasets, with over 110,079 images and 374 videos. These resources cover a variety of gastroenterological examination results, providing researchers with abundant data to train and test their algorithms. The richness of this dataset helps improve the application of machine learning models in real clinical settings.

The Kvasir-Capsule [12] dataset focuses on providing high-quality images from video capsule endoscopy. This dataset contains 47,238 annotated images and 117 videos, covering various anatomical features and pathological conditions of the digestive tract. The diversity and coverage of Kvasir-Capsule make it an important resource for capsule endoscopy research and machine learning applications.

Colon Polyp Segmentation and Detection Datasets. The CVC-12k (CVC-ClinicDB) [4], CVC-300 [13], and CVC-ColonDB [5] datasets are primarily used for polyp detection and segmentation in colonoscopy videos. CVC-12k (CVC-ClinicDB) provides 612 images, while CVC-300 and CVC-ColonDB offer 60 and 380 images, respectively. These datasets focus on the identification and image segmentation of colon polyps, providing critical visual information for early cancer screening.

The ETIS-Larib Polyp Database [11] is specifically designed for medical image analysis, with a focus on polyp detection in colonoscopy videos. This database contains 196 frames extracted from colonoscopy videos, each frame containing one or more instances of polyps. To aid in the development and evaluation of polyp detection algorithms, the database also includes ground truth

Table 1: The effectiveness of DataDEL. The table demonstrates that across nearly all datasets and various scales of pre-training data, the mIoU of pre-trained datasets is significantly higher compared to datasets that are directly fine-tuned using ImageNet parameters without pre-training. This indicates that pre-training with DataDEL can effectively enhance the performance of endoscopic downstream tasks.

Model	Dataset	mIoU (Pre-training Data)						
		5%	10%	20%	33%	50%	100%	ImageNet
ViT-Base	CVC-ClinicDB [4]	75.24	74.58	75.27	75.50	75.49	74.95	74.22
	CVC-ColonDB [5]	69.52	68.58	69.90	71.60	69.72	69.48	66.84
	CVC-300 [13]	63.67	56.69	61.40	62.85	61.97	61.16	40.72
	ETIS [11]	47.50	49.44	50.20	50.28	49.62	49.13	42.02
	ImageCLEFmed [7]	70.95	71.80	71.44	72.58	71.23	72.02	71.50
	Kvasir-Instrument [8]	79.38	79.52	80.22	80.38	80.48	79.70	79.27
	Kvasir-SEG [10]	75.45	75.74	76.37	76.03	76.77	76.02	75.43
	PolypGen2021 [2]	60.93	61.61	61.47	62.28	62.20	60.89	61.10

Table 2: MedDEL (DEL) time discrepancies in relation to remaining data and ratio across various similarity thresholds (η). This table illustrates the performance metrics of data processing at various similarity threshold values (η), encompassing details on remaining data, processing ratio, and the specific time spent on deduplication operations.

η	Remaining Data	Ratio	DEL Time
0.7	4,461	5%	3h:11 @4GPUS
0.75	9,055	10%	5h:28 @4GPUS
0.8	16,789	20%	8h:29 @4GPUS
0.85	28,352	33%	10h:59 @4GPUS
0.9	43,484	50%	11h:50 @4GPUS

Table 3: Training time discrepancies for MAE (ViT-Base) across different datasets and epochs. This table illustrates the training time for the MAE based on the ViT-Base model under varying proportions of pretraining data and different numbers of epochs. When the training data is reduced by a certain proportion, and simultaneously the number of epochs is increased by the same proportion, the training time remains nearly unchanged. Therefore, employing a smaller dataset and training for more epochs is fair in terms of training strategy.

Model	Dataset	epochs	Training Time
MAE (ViT-Base)	5% pretraining data	4,000	25h:20 @4GPUS
	10% pretraining data	2,000	28h:31 @4GPUS
	20% pretraining data	1,000	28h:52 @4GPUS
	33% pretraining data	600	27h:14 @4GPUS
	50% pretraining data	400	33h:04 @4GPUS

data of the polyps. This ground truth is in the form of masks corresponding to the areas covered by polyps in each image.

Endoscopy Artifact Detection and Disease Localization Datasets. The EAD2019 dataset [3] focuses on identifying and locating artifacts in endoscopic videos, a crucial step in developing effective computer-assisted diagnostic tools. Artifacts may include motion blur, changes in saturation, light reflections, etc., which can interfere with the accurate diagnosis of diseases. This dataset contains 375 unique video frames with artifacts, each annotated and validated by professional doctors. The dataset provides a rich sample of artifacts to help researchers more accurately analyze and identify diseased areas in complex endoscopic images. It includes 375 unique video frames with artifacts, each annotated and validated by professional doctors.

The EDD2020 [1] contains 380 annotated video frames specifically for disease detection and localization tasks. Each frame provides precise disease area annotations, manually completed by clinical experts, ensuring high quality and reliability. It offers a range of precisely annotated disease samples, useful for developing and evaluating endoscopic image analysis algorithms. Improvements in these algorithms will directly impact the accuracy and efficiency of clinical diagnoses, significantly contributing to the early detection and treatment of gastrointestinal diseases.

Other Specialized Medical Imaging Datasets. The Kvasir-SEG dataset [10] provides an important resource for the image segmentation of gastrointestinal polyps, containing 1000 endoscopic images of polyps and their corresponding ground truth segmentation masks. Kvasir-Sessile is a subset of this, including 196 images of striped, more challenging polyps and their annotations, offering more difficult samples for research.

The Kvasir-Instrument dataset [8] is a unique and valuable resource in the field of medical imaging and computer vision, particularly focusing on the segmentation of diagnostic and therapeutic tools used in gastrointestinal endoscopy. This dataset includes 590 annotated frames, featuring images of various gastrointestinal surgical tools such as sheaths, balloons, and biopsy forceps, along with their segmentation annotations. It is crucial for developing and testing algorithms for automatic recognition and segmentation of endoscopic tools.

Table 4: The training time of the ViT-Base model for downstream datasets corresponds to different proportions of pretraining data, training epochs, and batch size. This table illustrates the training outcomes of the ViT-Base model across various datasets and different proportions of pretraining data.

Model	Dataset	epochs	batch size	Training Time (Pre-training Data)					
				5%	10%	20%	33%	50%	100%
ViT-Base	CVC-ClinicDB	300	16	0h:52	0h:52	1h:15	1h:14	1h:16	1h:15
	CVC-ColonDB	500	16	1h:37	1h:37	1h:35	1h:39	1h:39	1h:36
	CVC-300	500	8	0h:38	0h:37	0h:40	0h:41	0h:43	0h:43
	ETIS	300	16	0h:42	0h:42	0h:42	0h:42	0h:43	0h:43
	ImageCLEFmed	200	16	0h:59	1h:01	1h:00	1h:01	0h:59	1h:08
	Kvasir-Instrument	200	16	0h:36	0h:52	0h:54	0h:52	0h:52	1h:01
	Kvasir-SEG	200	16	1h:25	1h:18	1h:18	1h:19	1h:21	1h:28
	PolypGen2021	500	16	5h:29	5h:38	5h:35	4h:27	5h:26	5h:40

ImageCLEFmed [7] provides a specialized collection of medical images for polyp segmentation. It contains 3,762 frames, covering polyps of different sizes and shapes, providing a rich resource for polyp identification and segmentation.

Sigrun L Eskeland, et al. 2021. Kvasir-Capsule, a video capsule endoscopy dataset. *Scientific Data* 8, 1 (2021), 142.

- [13] David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdal, Aaron Courville, et al. 2017. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering* 2017 (2017).

REFERENCES

[1] Sharib Ali, Noha Ghatwary, Barbara Braden, Dominique Lamarque, Adam Bailey, Stefano Realdon, Renato Cannizzaro, Jens Rittscher, Christian Daul, and James East. 2020. Endoscopy disease detection challenge 2020. *arXiv preprint arXiv:2003.03376* (2020).

[2] Sharib Ali, Debesh Jha, Noha Ghatwary, Stefano Realdon, Renato Cannizzaro, Osama E Salem, Dominique Lamarque, Christian Daul, Michael A Riegler, Kim V Anonsen, et al. 2021. PolypGen: A multi-center polyp detection and segmentation dataset for generalisability assessment. *arXiv preprint arXiv:2106.04463* (2021).

[3] Sharib Ali, Felix Zhou, Christian Daul, Barbara Braden, Adam Bailey, Stefano Realdon, James East, Georges Wagnieres, Victor Loschenov, Enrico Grisan, et al. 2019. Endoscopy artifact detection (EAD 2019) challenge dataset. *arXiv preprint arXiv:1905.03209* (2019).

[4] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilarino. 2015. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics* 43 (2015), 99–111.

[5] Jorge Bernal, Javier Sánchez, and Fernando Vilarino. 2012. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition* 45, 9 (2012), 3166–3182.

[6] 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. 2020. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific data* 7, 1 (2020), 283.

[7] William Hersh, Henning Müller, and Jayashree Kalpathy-Cramer. 2009. The ImageCLEFmed medical image retrieval task test collection. *Journal of Digital Imaging* 22 (2009), 648–655.

[8] Debesh Jha, Sharib Ali, Krister Emanuelsen, Steven A Hicks, Vajira Thambawita, Enrique Garcia-Ceja, Michael A Riegler, Thomas de Lange, Peter T Schmidt, Håvard D Johansen, et al. 2021. Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy. In *MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II* 27. Springer, 218–229.

[9] Debesh Jha, Vanshali Sharma, Neethi Dasu, Nikhil Kumar Tomar, Steven Hicks, MK Bhuyan, Pradip K Das, Michael A Riegler, Pål Halvorsen, Ulas Bagci, et al. 2023. GastroVision: A Multi-class Endoscopy Image Dataset for Computer Aided Gastrointestinal Disease Detection. In *Workshop on Machine Learning for Multi-modal Healthcare Data*. Springer, 125–140.

[10] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. 2020. Kvasir-seg: A segmented polyp dataset. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II* 26. Springer, 451–462.

[11] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. 2014. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery* 9 (2014), 283–293.

[12] Pia H Smedsrud, Vajira Thambawita, Steven A Hicks, Henrik Gjestang, Oda Olsen Nedrejord, Espen Næss, Hanna Borgli, Debesh Jha, Tor Jan Derek Berstad,