Endo-CLIP: Progressive Self-Supervised Pre-training on Raw Colonoscopy Records

Yili He[1,2,3]†, Yan Zhu[4,5]†, Peiyao Fu[4,5], Ruijie Yang[6], Tianyi Chen[1,3], Zhihua Wang[6], Quanlin Li[4,5], Pinghong Zhou[4,5], Xian Yang[7,8]✉, and Shuo Wang[1,3,5,8]✉

[1]Digital Medical Research Center, School of Basic Medical Sciences, Fudan University, Shanghai, China
[2]University College London, London, UK
[3]Shanghai Key Laboratory of MICCAI, Shanghai, China
[4]Endoscopy Center and Endoscopy Research Institute, Zhongshan Hospital, Fudan University, Shanghai, China
[5]Shanghai Collaborative Innovation Center of Endoscopy, Shanghai, China
[6]Shanghai Institute for Advanced Study of Zhejiang University, Shanghai, China
[7]Alliance Manchester Business School, The University of Manchester, Manchester, UK
[8]Data Science Institute, Imperial College London, London, UK

Abstract. Pre-training on image-text colonoscopy records offers substantial potential for improving endoscopic image analysis, but faces challenges including non-informative background images, complex medical terminology, and ambiguous multi-lesion descriptions. We introduce Endo-CLIP, a novel self-supervised framework that enhances Contrastive Language-Image Pre-training (CLIP) for this domain. Endo-CLIP's three-stage framework—cleansing, attunement, and unification—addresses these challenges by: (1) removing background frames, (2) leveraging large language models (LLMs) to extract clinical attributes for fine-grained contrastive learning, and (3) employing patient-level cross-attention to resolve multi-polyp ambiguities. Extensive experiments demonstrate that Endo-CLIP significantly outperforms state-of-the-art pre-training methods in zero-shot and few-shot polyp detection and classification, paving the way for more accurate and clinically relevant endoscopic analysis. Code will be made publicly available on `https://github.com/chrlott/EndoCLIP`.

Keywords: Language-Image Pre-training · Foundation Model · Endoscopy Image.

1 Introduction

Colorectal cancer remains a leading cause of cancer-related deaths globally [3]. Early detection and removal of precancerous polyps through colonoscopy significantly improves patient outcomes. Endoscopic examinations are crucial for

† Equal contribution.

✉ Corresponding authors: shuowang@fudan.edu.cn and xian.yang@manchester.ac.uk

this process, allowing clinicians to visualize the gastrointestinal tract and identify suspicious lesions. Computer-assisted systems have demonstrated significant improvements in adenoma detection rates (ADR) [15]. These systems primarily target two key tasks [2]: polyp detection (identifying the presence of polyps) and optical biopsy (determining polyp histology in real-time during the procedure). While artificial intelligence (AI)-driven models have achieved near-expert performance, their generalization and robustness across diverse clinical settings and patient populations remain limited [1], impeding adoption and clinical use.

Pre-training, a technique that learns generalizable visual representations from large datasets, has shown considerable promise in medical imaging [29,4]. By pre-training image encoders on massive datasets, subsequent transfer learning to specific downstream tasks becomes more effective, even with limited labeled data [9]. The vast, daily-accumulating archives of electronic colonoscopy records – comprising numerous screenshots and free-text reports – represent a largely untapped resource for such pre-training. Self-supervised learning, particularly contrastive learning approaches like DINO [6], offers a powerful means to leverage this data without manual annotations [12,19,31]. Contrastive Language-Image Pre-training (CLIP) [21] and its variants have demonstrated impressive success by aligning images and text in a shared embedding space. In the medical domain, adaptations like MedCLIP [27] have successfully leveraged image-text pairings in clinical settings, notably in X-ray diagnosis [30,22,16,7], ultrasound [8] and digital pathology [28,14].

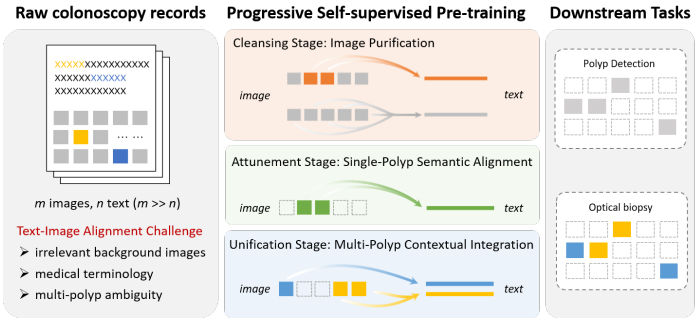

Fig. 1. Challenges of pre-training on colonoscopy records and the Endo-CLIP solution.

Although initial efforts have been made [26,5], adapting CLIP to colonoscopy records presents unique and substantial challenges. The inherent complexities of colonoscopy data create a fundamental mismatch between images and text [25]. In routine clinical practice, a single diagnostic report corresponds to numerous endoscopic images (often around 100 per examination), yet only a small fraction depict clinically significant findings (polyps), while the vast majority show normal mucosa. This creates a critical "needle-in-a-haystack" problem: accurately associating the few relevant polyp images with the corresponding, often brief,

textual descriptions within a lengthy report. Furthermore, the considerable morphological variability of polyps – in size, shape, texture, and color – demands that the model effectively interpret nuanced medical terminology. Finally, the frequent occurrence of multiple polyps within a single examination (approximately 60% in our cohort) introduces overlapping visual cues and potentially ambiguous textual descriptions, complicating the establishment of precise one-to-one correspondences between individual polyp images and segments of the report. To address these unique challenges, we present Endo-CLIP, a novel self-supervised framework for image-text alignment in raw colonoscopy records (Fig. 1). The pretrained Endo-CLIP models serves as backbones for downstream tasks, achieving high accuracy and robustness in polyp detection and classification, even under zero-shot and few-shot conditions.

2 Method

2.1 Model Overview

Endo-CLIP (Fig. 2) is a progressive self-supervised learning framework for endoscopic image-text alignment, addressing challenges like absent image-text pairs, background dominance, and variable polyp descriptions in colonoscopy data. It comprises three stages: (1) Cleansing: Filters non-informative background frames using diagnostic sentences indicating polyp presence/absence. (2) Attunement: Enhances fine-grained polyp representation by enforcing semantic consistency between single-polyp images and their morphological attributes. (3) Unification: Resolves multi-polyp matching ambiguity via a cross-attention mechanism, dynamically associating polyp images with corresponding text descriptions and integrating contextual relationships.

2.2 Cleansing Stage: Image Purification

In the first stage, we identify polyp-bearing images from raw colonoscopy records, mitigating the noise introduced by normal bowel images and enhancing the quality of subsequent learning. We employ LLMs to preprocess each patient's report R_i, removing non-diagnostic content and extracting key descriptions for each polyp. Given N medical cases $\{X_i, R_i\}$, where $X_i \in \mathbb{R}^{K_i \times H \times W \times C}$ contains K_i images and R_i consists of a variable number of diagnostic sentences, we randomly sample one image I_i and encode it using a vision encoder E_v as $\mathbf{v}_i = \mathrm{E}_v(I_i)$. If R_i describes polyps, we use the standardized sentence `'This is a colon with polyps.'`; otherwise, `'This is a normal background.'`. The selected sentence S_i is then encoded using a text encoder E_t as $\mathbf{t}_i = \mathrm{E}_t(S_i)$.

To associate patient-level images with their corresponding textual descriptions, we employ a CLIP-style contrastive loss. We gather all $\mathbf{v}_i$ and $\mathbf{t}_i$ for a batch of N patients into matrices $\mathbf{V}, \mathbf{T} \in \mathbb{R}^{N \times d}$. Let $\mathrm{cosim}(\cdot, \cdot)$ denote the cosine similarity function applied row-wise, yielding two $N \times N$ matrices:

$$\mathbf{S}^{(v \to t)} = \mathrm{cosim}(\mathbf{V}, \mathbf{T}) \in \mathbb{R}^{N \times N}, \quad \mathbf{S}^{(t \to v)} = \mathrm{cosim}(\mathbf{T}, \mathbf{V}) \in \mathbb{R}^{N \times N}. \tag{1}$$

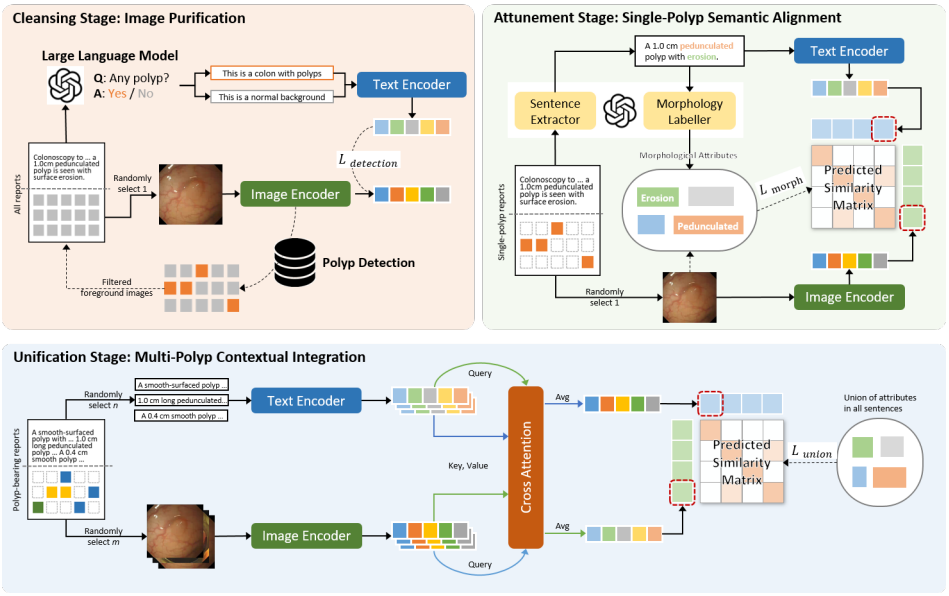

Fig. 2. Overview of Endo-CLIP: a progressive self-supervised pre-training framework that refines endoscopic image–text alignment.

Following the InfoNCE formulation [18], the detection loss is

$$\mathcal{L}_{\text{detection}} = \frac{1}{2}\Big[\text{CE}\big(\mathbf{S}^{(v\to t)}, \mathbf{Y}^{(v\to t)}\big) + \text{CE}\big(\mathbf{S}^{(t\to v)}, \mathbf{Y}^{(t\to v)}\big)\Big], \tag{2}$$

where $\text{CE}(\cdot,\cdot)$ is the cross-entropy loss computed on softmax-normalized similarities, and $\mathbf{Y}^{(v\to t)}, \mathbf{Y}^{(t\to v)}$ are one-hot labels indicating positive matches along the diagonal. Since each image I_i is randomly sampled, many polyp-positive cases may contain normal bowel regions, introducing noise into the initial training.

To refine detection, we filter polyp-bearing images using a prevalence-guided selection strategy [11]. Specifically, for polyp-positive patients, we compute the cosine similarity between each image and standardized polyp-positive and polyp-negative textual representations, followed by a softmax operation to obtain the probability of each image containing a polyp. Based on the known polyp prevalence, we retain the corresponding top percentage of highest-probability frames as polyp-bearing images. For polyp-negative patients, a randomly sampled frame remains as the negative example. We then reapply the loss in Eq. (2) using these filtered images, initializing from the first-round model. This two-round contrastive learning approach enhances the E_v as a polyp detector. The final outcome is a clean dataset of polyp-bearing images that sets the stage for fine-grained polyp representation learning in subsequent stages.

2.3 Attunement Stage: Single-Polyp Semantic Alignment

In the second stage, we focus on single-polyp cases to enhance fine-grained polyp representation learning by incorporating clinically defined morphological features. This approach mitigates false negatives arising from rigid one-hot supervision, which may overlook semantically similar pairs. We employ an LLM to extract morphological attributes of polyps from each diagnostic sentence S_i, following clinical guidelines. For the i-th single-polyp case, these attributes are encoded into two multi-hot vectors $\mathbf{a}_v^i$ and $\mathbf{a}_t^i$, ensuring semantic consistency between the polyp image I_i and its textual description S_i. Collecting these vectors for a batch of N single-polyp cases, we form two matrices $\mathbf{A}_v$ and $\mathbf{A}_t$, where row i corresponds to $(\mathbf{a}_v^i)^\top$ or $(\mathbf{a}_t^i)^\top$. From these attribute matrices, we compute the morphological similarity matrices $\mathbf{M}^{(v\to t)}$ and $\mathbf{M}^{(t\to v)}$, both in $\mathbb{R}^{N\times N}$. Each entry $\mathbf{M}_{ij}^{(v\to t)}$ is computed as the cosine similarity between $\mathbf{a}_v^i$ and $\mathbf{a}_t^j$, representing the morphological alignment between the i-th image attribute vector and the j-th text attribute vector.

Meanwhile, the model produces similarity matrices $\mathbf{S}^{(v\to t)}$ and $\mathbf{S}^{(t\to v)}$ for the same batch of N single-polyp cases, using the image–text embeddings described in Eq. (1). To integrate morphological knowledge, we replace the one-hot labels from Eq. (2) with the soft targets $\mathbf{M}^{(v\to t)}$ and $\mathbf{M}^{(t\to v)}$. Formally,

$$\mathcal{L}_{\text{morph}} = \frac{1}{2}\Big[\text{CE}\big(\mathbf{S}^{(v\to t)}, \mathbf{M}^{(v\to t)}\big) \;+\; \text{CE}\big(\mathbf{S}^{(t\to v)}, \mathbf{M}^{(t\to v)}\big)\Big]. \tag{3}$$

By incorporating morphological attributes into the contrastive objective, we reduce false negatives and achieve finer medical semantic alignment.

2.4 Unification Stage: Multi-Polyp Contextual Integration

In the third stage, we refine representation learning for multi-polyp cases, where lesion-text associations are ambiguous. We introduce a polyp-level cross-attention mechanism [23,24], allowing each image and text embedding to attend to all possible matches. The attended features are then aggregated at the patient level, enabling holistic alignment through semantic similarity learning.

Let $\{\mathbf{v}_i^1, \ldots, \mathbf{v}_i^{K_i}\}$ be the polyp image embeddings for patient i, where K_i is the number of images, and let $\{\mathbf{t}_i^1, \ldots, \mathbf{t}_i^{L_i}\}$ be the corresponding textual embeddings, where L_i is the number of descriptive sentences. Each polyp embedding $\mathbf{v}_i^k$ attends to all textual embeddings $\{\mathbf{t}_i^\ell\}$, and vice versa, using

$$\alpha_i^{k,\ell} = \text{softmax}\Big(\tfrac{(\mathbf{W}_Q\,\mathbf{v}_i^k)^\top(\mathbf{W}_K\,\mathbf{t}_i^\ell)}{\sqrt{d}}\Big), \quad \mathbf{o}_{i,k}^{(v\to t)} = \sum_{\ell=1}^{L_i} \alpha_i^{k,\ell}\,\big(\mathbf{W}_V\,\mathbf{t}_i^\ell\big), \tag{4}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are learnable projection matrices. Symmetrically, each textual embedding $\mathbf{t}_i^\ell$ attends over $\{\mathbf{v}_i^k\}$ to produce $\mathbf{o}_{i,\ell}^{(t\to v)}$. To form patient-level representations, we aggregate polyp-level features by averaging the updated image and text embeddings and then stack these aggregated embeddings from all

patients into matrices $\hat{\mathbf{V}}$ and $\hat{\mathbf{T}} \in \mathbb{R}^{N \times d}$, enabling batch-wise similarity computation.

To incorporate lesion-specific details, we unify the morphological attributes of all L_i polyps into a single multi-hot label vector, representing the full spectrum of lesion characteristics for each patient. This ensures that our patient-level contrastive learning is guided by comprehensive morphological supervision. We then construct the soft target matrix $\hat{\mathbf{M}}$.

We compute row-wise cosine similarities $\hat{\mathbf{S}}^{(v \to t)}$ and $\hat{\mathbf{S}}^{(t \to v)}$ with $\hat{\mathbf{V}}$ and $\hat{\mathbf{T}}$ according to Eq. (1), and define the Stage 3 loss as:

$$\mathcal{L}_{\text{union}} = \frac{1}{2}\Big[\text{CE}\big(\hat{\mathbf{S}}^{(v \to t)}, \hat{\mathbf{M}}^{(v \to t)}\big) + \text{CE}\big(\hat{\mathbf{S}}^{(t \to v)}, \hat{\mathbf{M}}^{(t \to v)}\big)\Big]. \tag{5}$$

By aggregating multi-polyp features at both the polyp and patient levels, we mitigate matching ambiguity in multi-lesion scenarios while maintaining the fine-grained semantic alignment from the Attunement Stage.

3 Experiment

3.1 Implementation Details

The vision encoder adopts the ViT-B/32 version of the Vision Transformer [10], while the text encoder utilizes a base Transformer architecture, both initialized with pre-trained CLIP weights. The model is optimized using AdamW with a learning rate of 5e-5. Training is conducted on eight NVIDIA A800 GPUs, with batch sizes of 32 for stage one and stage two, and 8 for stage three. Each stage is trained for 100 epochs, with the first 50 epochs dedicated to warm-up.
Pre-training Dataset: Our dataset comprises 13k colorectal examination cases with 874k endoscopic images and corresponding diagnostic reports. Polyp-positive cases account for 5.8k, including 2.5k multi-polyp cases (43% of positives).

3.2 Downstream Task Setup

Datasets and Evaluation Metrics: We contribute and publicly release EndoReport50, a dataset of 84 polyps from 50 patients with detailed lesion annotations, including textual descriptions, morphological attributes of polyps (nine clinically relevant aspects annotated by clinicians) and pathology reports, and conduct experiments on it to evaluate our model. We consider two classification tasks: 1) Polyp Detection, distinguishing normal bowel images from polyp-containing ones, using a highly imbalanced test set of 2,308 normal and 608 polyp images. 2) Polyp Malignancy Classification, where polyps are categorized as benign (Vienna = 1) or malignant (Vienna = 3,4,5). Each polyp is represented by its most clinically indicative image, yielding 16 benign and 68 malignant samples.

For both tasks, we report the Area Under the Receiver Operating Characteristic Curve (AUROC) and Area Under the Precision-Recall Curve (AUPR) to evaluate model performance under class imbalance. We compare models in

zero-shot and few-shot settings, with training data ratios of 5% and 10% for polyp detection, and 10% and 20% for malignancy classification.

Comparison Methods: We compare our method with baselines covering supervised, vision contrast, and vision-language contrastive learning. **Random** is a ResNet-50 [13] trained from scratch, while **ImageNet** is pretrained on ImageNet. **DINO** [6] and **SSCD** [20] use self-supervised learning, with SSCD tailored for medical imaging, and both of them pretrained on our dataset. **CLIP** [21] learns from large-scale image-text pairs for zero-shot transfer. All models are evaluated under the same setup for fair comparison.

3.3 Experimental Results

Comparison with baseline methods. Table 1 compares the performance of different methods on polyp detection and malignancy classification. Our Endo-CLIP models consistently outperform the baselines across all evaluation settings. For Polyp Detection, Endo-CLIP surpassing CLIP by +9.93% AUROC in zero-shot and maintaining the lead in few-shot scenarios. This highlights the effectiveness of our two-round contrastive filtering in Stage 1. Self-supervised models like DINO and SSCD lag behind, underscoring the benefits of multimodal pretraining. For Malignancy Classification, Endo-CLIP attains state-of-the-art performance with 92.13% AUROC in zero-shot, outperforming CLIP by +13.91%. Its advantage persists in 10% and 20% few-shot settings, confirming the impact of morphology-aware learning and cross-attention in Stages 2 and 3.

Table 1. Performance comparison of different methods on polyp detection and malignancy classification tasks. "-" indicates that the model does not support zero-shot learning. Endo-CLIP* refers to the model variant where the Cleansing Stage is used for polyp detection, while the Unification Stage is applied for malignancy classification.

Models	Polyp Detection						Malignancy Classification					
	AUROC			AUPR			AUROC			AUPR		
	Zero-shot	5%	10%	Zero-shot	5%	10%	Zero-shot	10%	20%	Zero-shot	10%	20%
Random	-	55.61	59.34	-	24.75	26.57	-	48.02	53.16	-	79.80	83.29
ImageNet	-	73.45	78.31	-	42.95	50.05	-	48.17	54.40	-	82.06	86.79
Dino	-	67.11	68.96	-	35.97	38.36	-	51.58	57.55	-	82.98	85.93
SSCD	-	57.07	62.85	-	27.15	31.91	-	52.12	53.41	-	83.51	84.96
CLIP	66.45	70.25	75.00	31.27	37.80	45.41	45.22	45.63	47.39	78.18	80.53	81.28
Endo-CLIP*	**76.38**	**79.08**	**82.38**	**45.60**	**50.38**	**54.37**	**72.79**	**58.72**	**62.73**	**92.13**	**87.68**	**89.61**

Feature Visualization. To gain deeper insights into the feature representations learned by different models, we utilize t-SNE [17] visualization to project high-dimensional feature embeddings into a two-dimensional space. Figure 3 illustrates the t-SNE projections of feature distributions for malignancy classification, comparing DINO, SSCD, CLIP, and Endo-CLIP. The DINO and SSCD models produce dispersed and overlapping clusters, suggesting their limited ability to distinguish between benign and malignant cases. CLIP, which benefits

from large-scale image-text pre-training, demonstrates improved separation, although some overlap between classes remains. In contrast, Endo-CLIP achieves the most distinct clustering, with well-defined boundaries between benign and malignant polyps. These results emphasize the effectiveness of hierarchical contrastive learning and morphology-aware supervision in refining polyp representations.

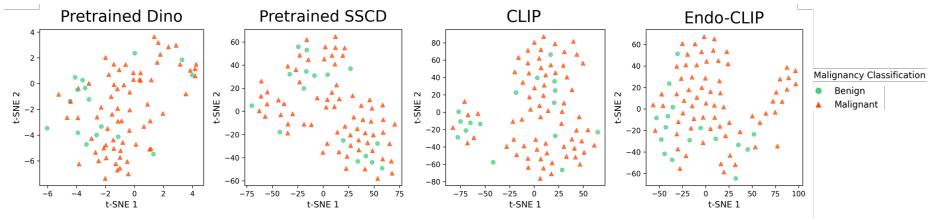

Fig. 3. t-SNE [17] visualization of feature distributions for malignancy classification. Each point represents a polyp image in the feature space, with green circles for benign cases and red triangles for malignant cases.

Ablation Study. Table 2 examines the impact of dataset composition and key methodological components on malignancy classification. Data ablation shows that incorporating both single-polyp (SP) and multi-polyp (MP) cases leads to improved performance, indicating that a diverse dataset aids generalization. On the methodological side, replacing Morphology Contrastive Learning (MC) with InfoNCE loss [18] results in degraded performance, suggesting that morphology-aware supervision provides useful semantic guidance. Similarly, replacing the Cross-Attention (CA) mechanism with simple averaging weakens multi-polyp feature integration, underscoring the role of explicit attention modeling. The full model achieves the highest scores, validating the effectiveness of our hierarchical framework in refining polyp representations.

Table 2. Results of ablation studies of Endo-CLIP.

Model	Data		Method		Malignancy Classification					
					AUROC			AUPR		
	SP	MP	MC	CA	Zero-shot	10%	20%	Zero-shot	10%	20%
Endo-CLIP					48.81	38.47	48.28	80.17	77.21	81.09
	✓				50.00	50.13	52.35	81.83	81.62	83.25
	✓		✓		39.71	54.23	59.36	78.25	84.78	87.32
	✓	✓	✓		48.35	58.05	55.17	84.33	84.76	86.50
	✓	✓	✓	✓	**72.79**	**58.72**	**62.73**	**92.13**	**87.68**	**89.61**

4 Conclusion

We introduced Endo-CLIP, a progressive self-supervised learning framework tailored for pre-training on raw colonoscopy records. To address the challenges of image-text alignment, we propose a three-stage approach: (1) Image purification filters out non-informative frames, (2) single-polyp semantic aligment leverages clinically defined attributes for fine-grained semantic learning, and (3) multi-polyp contextual integration applies cross-attention to resolve complex image-text associations at the patient level. Experimental results demonstrate that Endo-CLIP outperforms baselines in both zero-shot and few-shot settings for polyp detection and malignancy classification. By enhancing the alignment of endoscopic images and textual reports, Endo-CLIP lays a foundation for more accurate and clinically meaningful endoscopic analysis.

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

Acknowledgement. This study was supported in part by the National Key Research and Development Program of China (No. 2024YFF1207500), Shanghai Municipal Education Commission Project for Promoting Research Paradigm Reform and Empowering Disciplinary Advancement through Artificial Intelligence (SOF101020), the International Science and Technology Cooperation Program under the 2023 Shanghai Action Plan for Science (23410710400), the National Natural Science Foundation of China (82170555, 82203193), and the Shanghai Academic/Technology Research Leader (22XD1422400).

References

1. Ahmad, O.F., Mori, Y., Misawa, M., Kudo, S.e., Anderson, J.T., Bernal, J., Berzin, T.M., Bisschops, R., Byrne, M.F., Chen, P.J., et al.: Establishing key research questions for the implementation of artificial intelligence in colonoscopy: a modified delphi method. Endoscopy **53**(09), 893–901 (2021)
2. Ali, S.: Where do we stand in ai for endoscopic image analysis? deciphering gaps and future directions. npj Digital Medicine **5**(1), 184 (2022)
3. Arnold, M., Sierra, M.S., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global patterns and trends in colorectal cancer incidence and mortality. Gut **66**(4), 683–691 (2017)
4. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)
5. Batić, D., Holm, F., Özsoy, E., Czempiel, T., Navab, N.: Endovit: pretraining vision transformers on a large collection of endoscopic images. International Journal of Computer Assisted Radiology and Surgery **19**(6), 1085–1091 (2024)
6. 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)
7. Chen, X., He, Y., Xue, C., Ge, R., Li, S., Yang, G.: Knowledge boosting: Rethinking medical contrastive vision-language pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 405–415. Springer (2023)

8. Christensen, M., Vukadinovic, M., Yuan, N., Ouyang, D.: Vision–language foundation model for echocardiogram interpretation. Nature Medicine **30**(5), 1481–1488 (2024)
9. Dai, Y., Liu, F., Chen, W., Liu, Y., Shi, L., Liu, S., Zhou, Y., et al.: Swin mae: masked autoencoders for small datasets. Computers in biology and medicine **161**, 107037 (2023)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
11. Fang, A., Jose, A.M., Jain, A., Schmidt, L., Toshev, A., Shankar, V.: Data filtering networks. arXiv preprint arXiv:2309.17425 (2023)
12. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
14. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine **29**(9), 2307–2316 (2023)
15. Le Berre, C., Sandborn, W.J., Aridhi, S., Devignes, M.D., Fournier, L., Smaïl-Tabbone, M., Danese, S., Peyrin-Biroulet, L.: Application of artificial intelligence to gastroenterology and hepatology. Gastroenterology **158**(1), 76–94 (2020)
16. Lin, W., Zhao, Z., Zhang, X., Wu, C., Zhang, Y., Wang, Y., Xie, W.: Pmc-clip: Contrastive language-image pre-training using biomedical documents. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 525–536. Springer (2023)
17. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
18. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
19. 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)
20. Pizzi, E., Roy, S.D., Ravindra, S.N., Goyal, P., Douze, M.: A self-supervised descriptor for image copy detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14532–14542 (2022)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
22. Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P.: Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. Nature biomedical engineering **6**(12), 1399–1406 (2022)
23. 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)
24. Wang, F., Zhou, Y., Wang, S., Vardhanabhuti, V., Yu, L.: Multi-granularity cross-modal alignment for generalized medical visual representation learning. Advances in Neural Information Processing Systems **35**, 33536–33549 (2022)

25. Wang, S., Zhu, Y., Luo, X., Yang, Z., Zhang, Y., Fu, P., Wang, M., Song, Z., Li, Q., Zhou, P., et al.: Knowledge extraction and distillation from large-scale image-text colonoscopy records leveraging large language and vision models. arXiv preprint arXiv:2310.11173 (2023)
26. Wang, Z., Liu, C., Zhang, S., Dou, Q.: Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 101–111. Springer (2023)
27. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing. vol. 2022, p. 3876 (2022)
28. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al.: A whole-slide foundation model for digital pathology from real-world data. Nature **630**(8015), 181–188 (2024)
29. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
30. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine learning for healthcare conference. pp. 2–25. PMLR (2022)
31. Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P., et al.: A foundation model for generalizable disease detection from retinal images. Nature **622**(7981), 156–163 (2023)