Contrastive Learning under Noisy Temporal Self-Supervision for Colonoscopy Videos

Luca Parolari[1,2]✉, Pietro Gori[2], Lamberto Ballan[1], Carlo Biffi[3⋆], and Loic Le Folgoc[2⋆]

[1] Department of Mathematics, University of Padova, Padova, Italy
luca.parolari@phd.unipd.it
[2] LTCI, Telecom Paris, Institut Polytechnique de Paris, Palaiseau, France
[3] Cosmo Intelligent Medical Devices, Dublin, Ireland

Abstract. Learning robust representations of polyp tracklets is key to enabling multiple AI-assisted colonoscopy applications, from polyp characterization to automated reporting and retrieval. Supervised contrastive learning is an effective approach for learning such representations, but it typically relies on correct positive and negative definitions. Collecting these labels requires linking tracklets that depict the same underlying polyp entity throughout the video, which is costly and demands specialized clinical expertise. In this work, we leverage the sequential workflow of colonoscopy procedures to derive self-supervised associations from temporal structure. Since temporally derived associations are not guaranteed to be correct, we introduce a noise-aware contrastive loss to account for noisy associations. We demonstrate the effectiveness of the learned representations across multiple downstream tasks, including polyp retrieval and re-identification, size estimation, and histology classification. Our method outperforms prior self-supervised and supervised baselines, and matches or exceeds recent foundation models across all tasks, using a lightweight encoder trained on only 27 videos. Code is available at github.com/lparolari/ntssl.

Keywords: Contrastive Learning · Representation Learning · Temporal Self-Supervision · Colonoscopy Videos.

1 Introduction

Colonoscopy is the gold standard for colorectal cancer prevention. It enables the detection and removal of adenomatous polyps, *i.e.,* abnormal tissue growths in the colon that can become cancerous [11]. Computer-aided systems based on deep learning have demonstrated the potential to improve patient outcomes and the cost-effectiveness of screening [1]. In colonoscopy, deep learning systems support applications ranging from polyp detection to downstream tasks such as polyp classification and size estimation, and can further assist post-procedure

⋆ Shared senior authorship.

workflows including structured polyp reporting and lesion retrieval [23,25,27,33]. These tasks rely on robust representations of polyp *tracklets*, *i.e.,* sequences of consecutive detections of the same polyp, and typically depend on dense annotations. Obtaining such supervision requires localizing polyps throughout the procedure, forming tracklets and associating them with the underlying polyp entity. This process is costly, time-consuming, and needs specialized clinical expertise, limiting the scalability of supervised learning approaches. Self-supervised learning offers a promising direction to reduce reliance on such labeled data.

Prior work adopts the SimCLR framework [7] to learn polyp tracklet representations by splitting each tracklet into two disjoint segments and enforcing similarity between the resulting positive views [15]. Subsequent work aggregates tracklets through clustering to form a more stable polyp-level representation, improving robustness in downstream tasks [26]. Another approach leverages the momentum-contrast framework [12] and incorporates textual reports to learn multi-modal tracklet representations [33]. Recent work [25] exploits polyp entity annotations, *i.e.*, labels that link detections to the same underlying polyp, to build ground-truth positive associations between tracklets for supervised contrastive learning. With this explicit supervision, the method achieves superior performance over self-supervised alternatives, indicating a persistent gap between supervised and self-supervised learning.

In this work, we aim to narrow this gap while eliminating the reliance on polyp entity annotations. Colonoscopy presents a highly challenging visual environment for self-supervised learning. Semantic cues are often sparse, and polyp appearance can vary markedly over time due to changes in viewpoint, scale and illumination, tissue deformation, motion blur, occlusions, debris, and the presence of endoscopic instruments. Despite these difficulties, we observe that the clinical workflow remains structured and largely sequential. The endoscopist inspects the colon, removes a polyp when encountered, and then proceeds to the next one. We posit that this procedural structure arising from full-procedure videos provides a valuable self-supervisory signal. Observations captured close in time are likely to depict the same polyp entity. Thus, we build associations from temporal structure and employ contrastive learning to learn shared semantic information between such observations. However, the induced temporal associations may occasionally be incorrect, for instance when distinct polyps appear within a short time interval, leading to a noisy self-supervision signal. To address this, we propose a noise-aware loss that mitigates errors from wrong temporal associations, preventing them from degrading the learned representation.

We evaluate the effectiveness of the learned representations on a wide range of downstream tasks, including polyp retrieval and re-identification, size estimation, and histology classification–on open-access datasets including REAL-Colon [6], SUN [22], PolypSize [29] and PolypsSet [19]. Our method outperforms prior self-supervised as well as supervised baselines, highlighting the effectiveness of the proposed temporal self-supervision and noise-aware loss. We also compare against recent general-purpose and endoscopy-specific foundation models and

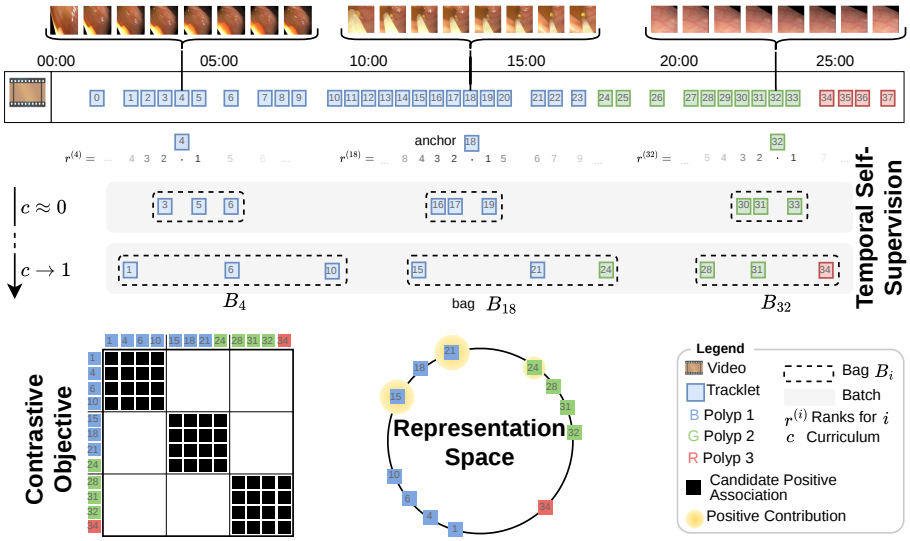

Fig. 1: Our method. (Top) Tracklets detected in a colonoscopy video are used to derive the self-supervision and produce bags of tracklets for contrastive learning. At the beginning of training ($c \approx 0$), bags are constituted by temporally adjacent tracklets, yielding low-variation views that are likely to correspond to the same polyp. Later ($c \to 1$), bags progressively include more distant tracklets, increasing diversity but also the chance of incorrect associations. (Bottom) The model learns from within-bag associations while limiting noisy ones. In the example, the loss maximizes the similarity between tracklet 18, 15 and 21 while down-weighting the contribution of noisy associations such as tracklet 24.

achieve comparable or superior results across all tasks using a lightweight architecture trained on only 27 videos.

In summary, this work focuses on learning robust polyp tracklet representations without relying on manual supervision. Our main contributions are threefold. (i) A self-supervised framework that exploits the sequential workflow of colonoscopy procedures to construct a self-supervision signal from temporal structure. (ii) A noise-aware contrastive loss that mitigates the impact of noisy temporal associations. (iii) An extensive evaluation on four public datasets, comparing to self-supervised, fully supervised and foundation model baselines.

2 Method

In this section, we present our framework for learning the representation of polyp tracklets under temporal self-supervision (Fig. 1). First, we describe how temporal structure can be used to construct a self-supervision signal. Then, we introduce our contrastive loss designed to learn from noisy associations.

2.1 Temporal Structure as Self-Supervision

Following related works [15,26,25], we start by constructing a dataset of N tracklets $\{x_i\}_{i=1}^{N}$ from ground-truth polyp detections, so that representation learning can be evaluated independently of detection and tracking errors. A tracklet is a sequence of fixed length L constituted by consecutive detections from the same polyp. Let v_i denote the colonoscopy video containing the i-th tracklet and p_i its temporal position within the video. Both v_i and p_i are inherently available from the video data and require no manual annotation. For each tracklet i, we define the same-video candidate set $\boldsymbol{\mathcal{C}} = \{j \in \{1, \dots, N\} \setminus \{i\} \mid v_j = v_i\}$ and compute an index $\pi_i(\cdot)$ to sort the C elements of $\boldsymbol{\mathcal{C}}$ by increasing temporal distance from i: $|p_{\pi_i(1)} - p_i| \leq |p_{\pi_i(2)} - p_i| \leq \cdots \leq |p_{\pi_i(C)} - p_i|$. The position in this ordering defines a ranking, where lower ranks correspond to tracklets that are temporally closer to i. Importantly, using ranks lets us define temporal neighborhoods without relying on absolute time windows or hand-tuned thresholds.
For each tracklet i, called *anchor*, our goal is to provide a set of candidate positive tracklets, called *bag*, for contrastive learning derived only from temporal associations.

Bag Construction Given K the bag size, $K \leq C$, a simple strategy is to build each bag by selecting the K samples that are closest in time to the anchor i and build the corresponding bag $B_i = \{x_{\pi_i(1)}, \dots, x_{\pi_i(K)}\}$. Although simple, this strategy yields highly redundant bags, making the contrastive task overly easy and limiting the diversity of positive views. We address this problem by sampling ranks from an exponential distribution. Specifically, given an anchor i, we sample K distinct ranks $r_\ell^{(i)} \geq 1, \ell \in \{1, \dots, K\}$ without replacement from: $\mathbb{P}_\tau(r) = \frac{\exp(-r/\tau)}{\sum_{u=1}^{C} \exp(-u/\tau)}$. The resulting bag is $B_i = \{x_m \mid m = \pi_i(r_\ell^{(i)}),\ \ell = 1, \dots, K\}$. The temperature τ shifts probability mass from lower to higher ranks, thereby controlling the diversity of tracklets in the bag. When $\tau \approx 0$, sampling concentrates on tracklets closest to i, yielding candidate positives that are highly similar to each other but very likely represent the same polyp. As τ increases, the distribution flattens and increasingly higher ranks are selected, resulting in candidate positives that capture larger temporal and appearance variation, but with a higher risk of including incorrect temporal associations (Fig. 1 (Top)).

Curriculum Rather than using a fixed temperature τ, we adopt curriculum learning [5] and gradually increase τ from low to high values over training. Specifically, we define a curriculum variable $c \in [0, 1]$ that tracks training progress in time, and set $\tau = \mathcal{T}(c)$ using a cosine schedule: $\mathcal{T}(c) = \tau_{\min} + \frac{1-\cos(\pi c)}{2}(\tau_{\max} - \tau_{\min})$. Fig. 2 (Left) visualizes the ranks sampled throughout the curriculum.

2.2 Noise-Aware Contrastive Loss

We consider a dataset of paired samples $\{(B_i, a_i)\}_{i=1}^{N}$, where a_i is an anchor tracklet and $B_i = \{b_{i1}, b_{i2}, \dots, b_{iK}\}$ is a bag of tracklets associated with a_i. The composition of B_i is crucial for effective contrastive representation learning.

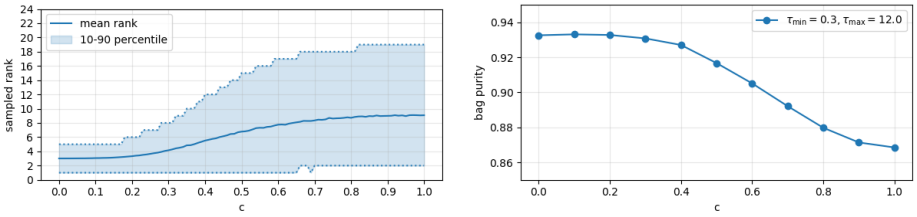

Fig. 2: (Left) Sampled rank throughout the curriculum. When c increases, higher ranks are sampled. (Right) Ratio of pure bags (*i.e.,* bags containing only tracklets from the same polyp) throughout the curriculum.

Positive samples should be sufficiently similar to the anchor to convey shared semantic content, yet diverse enough to promote robustness to appearance variations. Although colonoscopy procedures exhibit a sequential structure, temporal proximity does not always provide correct supervision. During the construction of bags, the sampler may include tracklets whose polyp entity differs from that of the anchor, particularly when higher-rank tracklets are considered. We refer to such bags as impure. Impure bags are undesirable because different polyps may exhibit different clinical properties (e.g. size or histology). Treating all tracklets in an impure bag as positives can introduce spurious supervision and degrade learning when applied naively within a contrastive framework. Fig. 2 (Right) illustrates this issue, showing that the proportion of impure bags increases as the curriculum progresses.

In light of this, we impose the following condition on the representation space that we want to learn: for each anchor a_i, at least one tracklet in B_i should be more similar to a_i than any tracklet in B_i is to any other anchor a_j. This condition relaxes traditional contrastive learning constraints where all positive samples contribute equally, such as in SupCon [18]. It can be satisfied by just one of the positives, allowing our model to ignore noisy associations and therefore prevent the degradation of the learned space [21] (Fig. 1 (Bottom)). Formally, let $f_\theta : x \mapsto f_\theta(x) \in \mathbb{R}^d$ be an encoder which maps tracklets into a shared representation space. Let $z_{ik} = f_\theta(b_{ik})$, $y_i = f_\theta(a_i)$ and $y_j = f_\theta(a_j)$. Let $s(\cdot,\cdot)$ be a similarity function in the joint embedding space. We can mathematically translate the condition into: $\max_k s(z_{ik}, y_i) > \max_k s(z_{ik}, y_j)$, $\forall j \neq i$. As commonly done in contrastive learning [3,10], this inequality can be transformed into an optimization problem by using the max operator and its smooth approximation log-sum-exp to derive the loss function:

$$\mathcal{L} = -\sum_i \log\left(\frac{\sum_{k=1}^{K}\exp\left(s(z_{ik}, y_i)\right)}{\sum_{k=1}^{K}\exp\left(s(z_{ik}, y_i)\right) + \sum_{\substack{j=1\\ j\neq i}}^{N}\sum_{k=1}^{K}\exp\left(s(z_{ik}, y_j)\right)}\right) \tag{1}$$

Multi-Level Objective The proposed loss is architecture-agnostic. To ensure fair comparison with prior work, we adopt the same lightweight tracklet encoder

used in related approaches [15,26,25]. It uses a transformer encoder with a learnable token [8] prepended to the sequence to aggregate a global representation of the tracklet. In addition, the encoder produces L embeddings, one per frame, analogous to patch tokens in Vision Transformers [9]: $\{z^{(t)}\}_{t=1}^{L}, z^{(t)} \in \mathbb{R}^d$. Our loss can be applied to the tracklet embedding, the per-frame embeddings, or both, by optimizing the sum of the corresponding objectives. We later present an ablation study to compare these variants.

3 Experiments

Datasets and Downstream Setup We train and evaluate on publicly available datasets, following the setup proposed in [26]. Specifically, we train on 27 videos (85 polyps) from REAL-Colon [6]. It contains 60 full-procedure recordings with frame-level bounding-box annotations and lesion-level metadata, *e.g.*, histology and size. We evaluate the generalization of the learned polyp tracklet representations on four downstream tasks.

Retrieval. This task measures the ability to retrieve same-polyp tracklets. Given a query tracklet, we rank all other tracklets by cosine similarity in the embedding space. We evaluate retrieval on REAL-Colon test set (19 videos, 47 polyps) and report performance using mean Average Precision (mAP) and Hit Rate@K (HR@K). A hit occurs if at least one tracklet with the same polyp identity is retrieved within the top-K results.

ReID. Re-identification measures the ability to decide whether two tracklets depict the same polyp entity or not. Using frozen embeddings, we compute the similarity between each embedding pair and threshold the scores to classify same-polyp vs different-polyp pairs. We employ the same evaluation set as in the retrieval task and measure performance with AUROC and AUPR.

Size estimation. It is formulated as binary classification (diminutive, *i.e.* $\leq$ 5mm, vs non-diminutive) using non-linear probing. We freeze the encoder and train a classifier. The evaluation set combines REAL-Colon [6] test set with two out-of-distribution datasets, namely SUN database [22] and PolypSize [29], totaling 161 videos (189 polyps). We split it by polyp entities, using 70% for training and 30% for evaluation. We report identity-weighted macro F1-Score.

Histology classification. It is treated as binary classification (adenoma vs non-adenoma) with the same non-linear probing setup. The evaluation set extends that of the size estimation task by additionally including PolypsSet [19], for a total of 195 videos (210 polyps). Performance is measured with accuracy.

Implementation Details The encoder is implemented following [25]. Frame embeddings are extracted with a ResNet-50 [13] and aggregated by a stack of three transformer encoder layers [30] with 8 heads, $d_{\text{model}} = 256$, $d_{\text{feedforward}} = 1024$ and dropout $= 0.1$. Following literature [7], we project the representations for contrastive learning with an MLP ($d_{\text{hidden}} = 256$, ReLU, $d_{\text{out}} = 128$). The bag size is set to $K = 4$, temperatures to $\tau_{\min} = 0.3$, $\tau_{\max} = 12.0$, batch size to 60. Following [26,25], a tracklet is formed from ground-truth polyp detections

Method	Params (M)	FLOPs (G)	*Retrieval* mAP (↑)	HR@1 (↑)	HR@5 (↑)	*ReID* AUROC (↑)	AUPR (↑)	*Size* F1-Score (↑)	*Histology* Acc (↑)
General Purpose FM									
DinoV2-S/14 [24]	22.1	44.2	47.31	89.81	95.52	80.07	29.77	62.81	61.68
DinoV2-B/14 [24]	86.6	175.7	46.40	87.77	95.79	78.74	26.99	67.97	66.57
DinoV2-L/14 [24]	304.4	622.5	43.08	87.50	94.97	75.64	23.77	68.02	70.63
DinoV2-g/14 [24]	1136	2331	48.29	**90.35**	95.92	78.23	27.04	**70.22**	70.14
DinoV3-S/16 [28]	21.6	34.7	48.95	88.99	<u>96.33</u>	80.47	31.77	64.57	67.13
DinoV3-B/16 [28]	85.7	137.7	47.17	88.45	96.06	77.04	26.54	<u>69.67</u>	71.54
DinoV3-L/16 [28]	303.1	487.2	43.89	87.91	94.97	68.97	20.54	66.58	<u>75.59</u>
V-JEPA2 [2]	326.0	272.1	29.01	70.79	88.32	69.84	17.14	68.68	61.54
Endoscopy FM									
EndoFM [31]	121.3	190.1	45.75	87.64	94.70	79.71	31.91	75.15*	68.32
EndoFM-LV [32]	121.3	190.1	28.21	63.18	81.39	67.56	19.28	61.99	68.60
EndoViT [4]	86.6	134.9	31.12	68.34	85.33	68.26	17.51	56.52	61.54
SurgeNet [16]	24.3	31.0	44.75	85.87	94.70	75.86	25.23	68.83	64.62
Fully supervised									
Sup [25]	25.6	33.1	54.72	81.93	94.97	93.85	41.58	62.35	72.66
TA [25]	25.6	33.1	<u>57.52</u>	84.24	95.24	**94.33**	43.81	63.87	73.29
Self-supervised									
TPC [26]	25.6	33.1	43.25	75.82	91.85	78.40	27.58	64.13	64.06
MVE [15]	25.6	33.1	42.05	81.25	94.16	79.73	27.47	58.82	66.22
SFE [15]	28.0	33.1	51.04	86.96	95.52	89.36	<u>46.13</u>	65.35	63.22
Ours	25.6	33.1	**63.13**	**90.22**	**96.47**	<u>94.17</u>	**51.94**	**70.24**	**82.38**

Table 1: Comparison with State-of-the-Art. All metrics are in percentage. **Best**, <u>second best</u>; shared bolding denotes no statistical difference ($p > 0.05$). *We note that EndoFM [31] is trained on SUN-SEG [17], a superset of SUN [22] annotated with segmentation masks, leading to an overlap with our test set.

whose Intersection Over Union is at least 0.1 between consecutive frames, subsampled to 1 frame every 4, cropped to 5× the diagonal of the bounding box, and resized to 232×232. Tracklet length is set to $L = 8$. For the non-linear probing setup we use an MLP ($d_{\text{hidden}} = 256$, GELU [14], dropout = 0.1), trained for 20 epochs using AdamW [20] with learning rate of 10^{-4} for size, 10^{-5} for histology. Batch size is 64. Hardware: 8 CPU cores, 64 GB RAM, NVIDIA A100 (64 GB).

Comparison with State-of-the-Art We compare our approach to state-of-the-art methods in Tab. 1. For fair comparison, fully supervised and self-supervised baselines are re-trained on the same data used in our experiments. Our method outperforms all prior self-supervised approaches, with 23.69% relative improvement on retrieval mAP, 5.38% and 12.60% on ReID AUROC and AUPR, 7.48% on size F1-Score and 24.4% on histology accuracy. This large margin highlights the effectiveness of the proposed temporal self-supervision and noise-aware loss. Compared to fully supervised baselines, our method consistently achieves superior performance across tasks. On ReID, AUROC is comparable to TA [25], while AUPR improves by 18.56%, indicating higher precision (fewer false positive matches) at comparable recall. For completeness, we also report results for both general-purpose and endoscopy-specific foundation mod-

		Retrieval			*ReID*		*Size*	*Histology*
		mAP	HR@1	HR@5	AUROC	AUPR	F1-Score	Acc
Exp. Sampling	Curriculum							
✗	✗	54.92	89.40	**97.28**	91.92	50.53	69.10	80.07
✓	✗	58.83	89.67	96.60	92.77	48.23	66.43	80.63
✓	✓	**63.13**	**90.22**	96.47	**94.17**	**51.94**	**70.24**	**82.38**
Frame	Tracklet							
✓	✗	62.15	90.90	97.42	93.85	50.84	66.83	80.42
✗	✓	61.49	**91.17**	**97.55**	93.89	51.78	**70.20**	81.82
✓	✓	**63.13**	90.22	96.47	**94.17**	**51.94**	**70.24**	**82.38**

Table 2: Ablation study. (Top) Impact of the exponential rank-based sampling used to construct bags, and curriculum strategy that controls the trade-off between safer and more diverse associations during training. (Bottom) Effect of applying the proposed loss on frame or tracklet embeddings, or both.

els. The Dino family provides a strong baseline, yet our method performs better overall; we match or surpass DinoV2-giant on all tasks while using less than 3% of its parameters and being trained on only 27 videos, whereas DinoV2 models are pretrained on hundreds of millions of images. This improved efficiency (fewer parameters and FLOPs) also facilitates deployment, making in-procedure support more practical. Endoscopy-specific foundation models generally underperform general-purpose ones, consistent with their smaller scale and training data biased toward surgical videos rather than colonoscopy.

Ablation Study In Tab. 2 (Top) we analyze the impact of the exponential sampling over ranks and curriculum. We observe that even without exponential sampling or curriculum, our approach already surpasses the prior self-supervised baselines [15,26], underscoring the effectiveness of the noise-aware loss at leveraging temporal supervision. Adding these components yields consistent improvements, with the largest gains on retrieval and re-identification. On retrieval, despite comparable HR@1 and HR@5, the 14.95% mAP increase indicates improved global ranking and a more coherent embedding space. In Tab. 2 (Bottom) we study the effect of applying the loss at the tracklet level, frame level, or both. The latter yields the best overall performance. On retrieval, hit rate is slightly higher with tracklet-only embeddings, likely because they more directly emphasizes top-ranked matches rather than overall ranking consistency. Frame embeddings alone consistently underperform, suggesting noisier supervision. Nevertheless, all versions outperform self-supervised and supervised baselines.

4 Conclusion

We present a self-supervised framework for tracklet representation learning in colonoscopy. Self-supervision is derived from temporal structure and optimized with a contrastive loss that mitigates noisy associations. Extensive experiments

demonstrate strong transfer across downstream tasks, outperforming prior self-supervised and supervised baselines and matching or exceeding foundation models with a lightweight encoder while being trained on only 27 videos. These findings enable scalable representation learning in colonoscopy, reducing reliance on annotations while maintaining performance on clinical applications.

Acknowledgments. We acknowledge ISCRA for awarding this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy).

Disclosure of Interests. C.B. is affiliated with Cosmo Intelligent Medical Devices, the developer of the GI Genius medical device.

References

1. Areia, M., Mori, Y., Correale, L., Repici, A., Bretthauer, M., Sharma, P., Taveira, F., et al.: Cost-effectiveness of artificial intelligence for screening colonoscopy: A modelling study. The Lancet Digital Health **4**(6), e436–e444 (2022)
2. Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Komeili, M., Muckley, M.J., Rizvi, A., et al.: V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985 (2025)
3. Barbano, C.A., Dufumier, B., Tartaglione, E., Grangetto, M., Gori, P.: Unbiased supervised contrastive learning. In: ICLR (2023)
4. Batic, 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)
5. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: International Conference on Machine Learning. pp. 41–48 (2009)
6. Biffi, C., Antonelli, G., Bernhofer, S., Hassan, C., Hirata, D., Iwatate, M., Maieron, A., Salvagnini, P., Cherubini, A.: REAL-Colon: A dataset for developing real-world AI applications in colonoscopy. Scientific Data **11**(1), 539 (2024)
7. 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 (2020)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Association for Computational Linguistics. pp. 4171–4186 (2019)
9. 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)
10. Dufumier, B., Barbano, C.A., Louiset, R., Duchesnay, E., Gori, P.: Integrating Prior Knowledge in Contrastive Learning with Kernel. In: ICML (2023)
11. Hassan, C., Spadaccini, M., Iannone, A., Maselli, R., Jovani, M., Chandrasekar, V.T., Antonelli, G., Yu, H., Areia, M., et al.: Performance of artificial intelligence in colonoscopy for adenoma and polyp detection: A systematic review and meta-analysis. Gastrointestinal Endoscopy **93**(1), 77–85 (2021)

12. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: 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: Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
14. Hendrycks, D., Gimpel, K.: Gaussian error linear units. In: arXiv preprint arXiv:1606.08415 (2016)
15. Intrator, Y., Aizenberg, N., Livne, A., Rivlin, E., Goldenberg, R.: Self-supervised polyp re-identification in colonoscopy. In: Medical Image Computing and Computer Assisted Intervention. pp. 590–600 (2023)
16. Jaspers, T.J.M., de Jong, R.L.P.D., Li, Y., Kusters, C.H.J., Bakker, F.H.A., van Jaarsveld, R.C., Kuiper, G.M., van Hillegersberg, R., Ruurda, J.P., Brinkman, W.M., Pluim, J.P.W., et al.: Scaling up self-supervised learning for improved surgical foundation models. Medical Image Analysis **108**, 103873 (2026)
17. Ji, G.P., Xiao, G., Chou, Y.C., Fan, D.P., Zhao, K., Chen, G., Gool, L.V.: Video polyp segmentation: A deep learning perspective. International Journal of Automation and Computing **19**(6), 531–549 (2022)
18. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: Advances in Neural Information Processing Systems (2020)
19. Li, K., Fathan, M.I., Patel, K., Zhang, T., Zhong, C., Bansal, A., Rastogi, A., Wang, J.S., Wang, G.: Colonoscopy polyp detection and classification: Dataset creation and comparative evaluations. PLoS One **16**(8), e0255809 (2021)
20. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
21. Miech, A., Alayrac, J.B., Smaira, L., Laptev, I., Sivic, J., Zisserman, A.: End-to-end learning of visual representations from uncurated instructional videos. In: Conference on Computer Vision and Pattern Recognition. pp. 9876–9886 (2020)
22. Misawa, M., ei Kudo, S., Mori, Y., Hotta, K., Ohtsuka, K., Matsuda, T., Saito, S., Kudo, T., Baba, T., Ishida, F., et al.: Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database. Gastrointestinal Endoscopy **93**(4), 960–967 (2021)
23. Nogueira-Rodríguez, A., Domínguez-Carbajales, R., López-Fernández, H., Águeda Iglesias, Cubiella, J., Fdez-Riverola, F., Reboiro-Jato, M., Glez-Peña, D.: Deep neural networks approaches for detecting and classifying colorectal polyps. Neurocomputing **423**, 721–734 (2021)
24. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., et al.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research **2024** (2024)
25. Parolari, L., Cherubini, A., Ballan, L., Biffi, C.: Temporally-aware supervised contrastive learning for polyp counting in colonoscopy. In: Medical Image Computing and Computer Assisted Intervention (2025)
26. Parolari, L., Cherubini, A., Ballan, L., Biffi, C.: Towards polyp counting in full-procedure colonoscopy videos. In: IEEE International Symposium on Biomedical Imaging. pp. 1–5 (2025)
27. Sekiguchi, M., Mizuguchi, Y., Kawagoe, R., Saito, Y.: Artificial intelligence and its impact on the quality of endoscopy reports. Digestive Endoscopy **38**(1), e70057 (2026)

28. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S.E., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
29. Song, Y., Du, S., Wang, R., Liu, F., Lin, X., Chen, J., Li, Z., Li, Z., Yang, L., Zhang, Z., Yan, H., Zhang, Q., Qian, D., Li, X.: Polyp-Size: A precise endoscopic dataset for AI-driven polyp sizing. Scientific Data **12**(1), 918 (2025)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Łukasz Kaiser, Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)
31. Wang, Z., Liu, C., Zhang, S., Dou, Q.: Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In: Medical Image Computing and Computer Assisted Intervention. pp. 101–111 (2023)
32. Wang, Z., Liu, C., Zhu, L., Wang, T., Zhang, S., Dou, Q.: Improving foundation model for endoscopy video analysis via representation learning on long sequences. IEEE Journal of Biomedical and Health Informatics **29**(5), 3526–3536 (2025)
33. Xiang, S., Liu, C., Ruan, J., Cai, S., Du, S., Qian, D.: VT-ReID: Learning discriminative visual-text representation for polyp re-identification. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 3170–3174 (2024)