# Temporally-Aware Supervised Contrastive Learning for Polyp Counting in Colonoscopy

Luca Parolari[1]✉[0000−0001−8574−4997], Andrea Cherubini[2,3][0000−0002−5946−4390], Lamberto Ballan[1][0000−0003−0819−851X], and Carlo Biffi[2][0000−0002−4913−7441]

[1] Department of Mathematics, University of Padova, Padova, Italy
luca.parolari@phd.unipd.it
[2] Cosmo Intelligent Medical Devices, Dublin, Ireland
[3] Milan Center for Neuroscience, University of Milano–Bicocca, Milan, Italy

**Abstract.** Automated polyp counting in colonoscopy is a crucial step toward automated procedure reporting and quality control, aiming to enhance the cost-effectiveness of colonoscopy screening. Counting polyps in a procedure involves detecting and tracking polyps, and then clustering tracklets that belong to the same polyp entity. Existing methods for polyp counting rely on self-supervised learning and primarily leverage visual appearance, neglecting temporal relationships in both tracklet feature learning and clustering stages. In this work, we introduce a paradigm shift by proposing a supervised contrastive loss that incorporates temporally-aware soft targets. Our approach captures intra-polyp variability while preserving inter-polyp discriminability, leading to more robust clustering. Additionally, we improve tracklet clustering by integrating a temporal adjacency constraint, reducing false positive re-associations between visually similar but temporally distant tracklets. We train and validate our method on publicly available datasets and evaluate its performance with a leave-one-out cross-validation strategy. Results demonstrate a 2.2x reduction in fragmentation rate compared to prior approaches. Our results highlight the importance of temporal awareness in polyp counting, establishing a new state-of-the-art. Code is available at github.com/lparolari/temporally-aware-polyp-counting.

**Keywords:** Colonoscopy · Polyp Counting · Computer-Aided Diagnosis.

## 1 Introduction

Polyp counting in colonoscopy aims to determine the number of distinct polyps observed during the entire procedure. It involves detecting polyps in video frames, tracking their appearances over time to form tracklets—sequences of consecutive detections, and associating together tracklets that belong to the same polyp entity [15,24]. Once associations are established, the polyp count can be derived, enabling the computation of quality metrics such as Polyps Per Colonoscopy (PPC)

or, by adding polyp classification, the Adenoma Detection Rate (ADR). Successfully addressing this challenge would facilitate automated report generation, enhance cost-efficiency, and support endoscopist skills assessment [10,1,19,27].

In contrast to polyp detection and tracking [23,3,22,2], the task of polyp counting has received limited attention, with only two studies published to date [15,24]. These studies assume tracklets as given and focus on learning a robust visual representation for clustering. In [15], authors propose a contrastive learning approach based on the SimCLR framework [6]. They generate positive samples by automatically splitting tracklets into segments to create two "views" of the same polyp. In [24], authors refine this approach by integrating curriculum learning that gradually integrates harder samples, and introduce Affinity Propagation [9] clustering to enhance polyp counting. Recent studies [5,30] have investigated polyp retrieval through self-supervision and contrastive learning, defining the task as matching a query polyp against a large gallery of candidates. Unfortunately, this research has been limited to a private dataset comprising multiple colonoscopies from the same patients.

We note that all previous works leverage self-supervision for representation learning. While learning object semantics through positive views from data augmentation has proven highly effective for standard imaging scenario [13,12,4,11], its application in colonoscopy is hindered by the severe variations in both inter- and intra-polyp appearance due to motion blur, lighting, occlusions, camera distance, debris, and intermittent visibility of polyps. In this paper, we make a two-fold contribution to overcome this issue.

First, we learn both intra-polyp invariance and inter-polyp separability leveraging polyp entity information. Specifically, we adopt a supervised contrastive loss [16] where multiple positives samples are considered. This loss allows to explicitly model sample's relationships in the embedding space and promotes stronger invariance among views of the same polyp, while still ensuring sufficient separability. Additionally, using multiple positive samples reduces the reliance on complex hard positive or negative mining strategies [28], such as the curriculum learning approach adopted in previous work [24].

Second, we integrate the temporal information intrinsic in video data [25]. By leveraging temporal coherence, we aim to enhance intra-polyp invariance. Specifically, we hypothesize that enforcing similarity between temporally close tracklets, in particular for those exhibiting visual dissimilarity, will refine the learned embedding space. Inspired by soft-contrastive learning on natural images [26,17], we modify the loss function to weight targets based on temporal distance. Moreover, we further integrate temporal awareness in the clustering stage as a penalty term. The objective is to penalize associations between visually similar tracklets that are unlikely to belong to the same polyp due to their large temporal gap.

We conduct an extensive experimental investigation on public datasets. While [15] uses a private dataset, [24] proposes train, validation and test splits on the publicly available REAL-Colon dataset. Due to limited availability of data we introduce two changes. First, we expand the proposed training set with three

open-access datasets: LDPolyp [20], SUN [21] and PolypSet [18], totalying 384 polyps (299 more than in previously proposed split). For fair comparison, we retrain previous methods. Then, we transition from a simple validation and test split to a thoughtful evaluation using leave-one-out cross-validation (LOOCV). Results demonstrate that our method is the state-of-the-art in polyp counting obtaining an improvement of 2.2x with respect to previous works.

In summary, this work focuses on tracklet representation learning and clustering for polyp counting. Our main contributions include: (1) a shift from self-supervised to supervised contrastive learning, improving inter-polyp separability; (2) a contrastive loss with temporally-aware soft targets for better intra-polyp invariance; (3) a tracklet clustering method using temporal penalties to discourage unlikely associations; and (4) a comprehensive benchmark against competing methods, leveraging an expanded dataset and a robust LOOCV evaluation.

## 2 Method

Our approach is based on two main steps. First, the visual-temporal encoder, depicted in Fig. 1 (Top), obtains tracklet representation. We train the encoder with a novel temporally-aware supervised contrastive loss, where soft-targets are introduced to reflect temporal proximity between tracklets. This encourages embeddings of temporally adjacent but visually dissimilar samples to be considered similar while maintaining class discriminability. Then, we employ a clustering module to re-associate tracklets into polyp entities and count them. We represent this module in Fig. 1 (Bottom). This step leverages both the visual features extracted by the encoder and temporal information derived from the video. Specifically, we integrate a temporal penalty term to penalize the association of distant tracklets, thereby reducing false positive re-associations that occur between temporally distant tracklets in the video.

### 2.1 Learning temporally-aware polyp representation

The visual-temporal encoder is responsible for generating tracklets representation, and is depicted in Fig. 1 (Top). Its architecture is based on [15] and involves two components: a frame-level feature extractor and sequence-level aggregator. The frame-level feature extraction is implemented by a visual backbone, e.g. ResNet [14], to capture the appearance of polyps in individual frames. The frame embeddings are then processed by a transformer network [29] to capture frame dependencies. The final tracklet representation is obtained from the classification token (CLS) embedding [8], and projected by a non-linear head [6]. Specifically, the encoder receive a fragment $\boldsymbol{f}_i = \boldsymbol{t}_{[i\cdot\kappa:i\cdot\kappa+\kappa]}, i \in [0,\dots,M]$ and returns its embedding $\boldsymbol{e}_i \in \Re^d$. Here, $\boldsymbol{t}$ is a tracklet from video $\mathcal{V}$, and $M = \lfloor N/\kappa \rfloor$, with $N$ the actual length of the tracklet. Thus, each fragment encodes $\kappa$ images $\boldsymbol{f}_i = [\boldsymbol{x}_i^1,\dots,\boldsymbol{x}_i^\kappa]$. Each image $\boldsymbol{x}_i^j$ represents a crop to the bounding box of the polyp instance. To include context around the polyp, we enlarge the bounding box to $\psi$ times its diagonal before cropping.

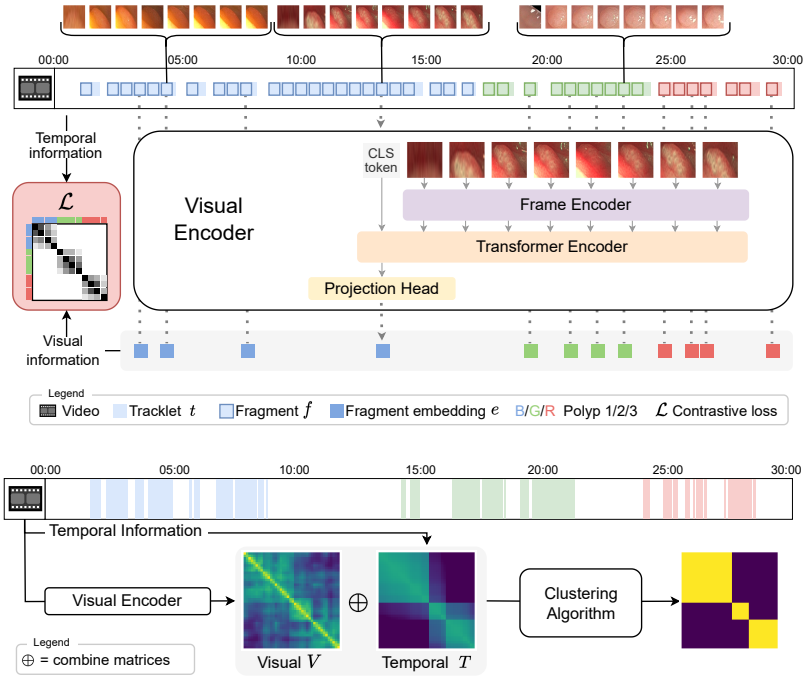


Fig. 1: Our method. *(Top)* Polyp tracklets detected in a colonoscopy video are processed by the encoder to extract their visual-temporal feature representation. The encoder is trained with a supervised contrastive loss using temporally-aware soft targets, computed on batches that include multiple views of multiple polyps, potentially from different videos. *(Bottom)* The clustering module obtains the tracklet embeddings and computes the similarity matrix. This matrix is combined with the temporal adjacency matrix, and a clustering algorithm derives the set of polyp entities.

We train the encoder with the contrastive loss defined in [16], which extends the normalized temperature-scaled cross-entropy loss used in SimCLR [6]. This loss considers multiple positives per anchor, rather than just a single one, allowing the model to explicitly define inter-sample relationships in the embedding space. Instead of equally representing all positive views in the ground truth distribution, we weight each view by its temporal distance with respect to the anchor. We depict our approach compared to traditional self-supervised and supervised in Fig. 2. This approach promotes intra-polyp consistency without sacrificing the ability to distinguish between different polyp instances and reduces sensitivity to appearance variation.

Formally, we formulate the loss as a matching problem [28], where an encoded anchor sample $\boldsymbol{a}$ is compared against a set of candidates $\{\boldsymbol{b}_1, \boldsymbol{b}_2, ..., \boldsymbol{b}_K\}$. The probability of $\boldsymbol{a}$ matching each candidate $\boldsymbol{b}_i$ is modeled using a softmax

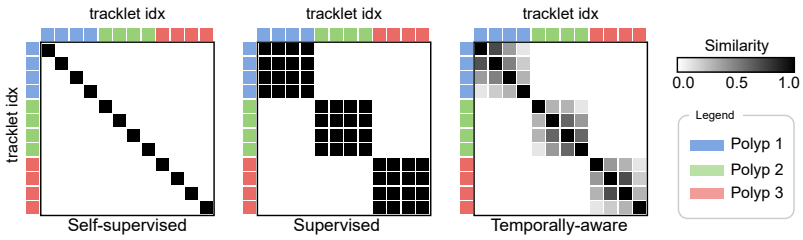


Fig. 2: Comparison of similarity adjacency matrices: self-supervised (left), supervised (center), and our temporally-aware contrastive loss (right). We visualize pairwise similarities for 12 tracklets from 3 polyps. Self-supervised learning does not explicitly define relationships between samples, supervised learning exploits polyp entity information to improve inter-polyp features modeling, while our approach improves also intra-polyp features by integrating temporal information.

distribution:

$$\boldsymbol{q}_i = \frac{\exp(\boldsymbol{a} \cdot \boldsymbol{b}_i/\tau)}{\sum_{j=1}^{K} \exp(\boldsymbol{a} \cdot \boldsymbol{b}_j/\tau)} \tag{1}$$

where $\tau$ is a temperature parameter, and all vectors have been $\ell_2$ normalized. To introduce temporal awareness in the ground-truth distribution, we obtain a vector $\boldsymbol{d}$ that encodes the temporal distances between $\boldsymbol{a}$ and all $\boldsymbol{b}_i$, i.e. $\boldsymbol{d} = |\boldsymbol{\pi} - \pi|/P$ where $\pi$ and $\boldsymbol{\pi}$ represent the timestamps of $\boldsymbol{a}$ and all $\boldsymbol{b}_i$ in the video, $P$ is the length of the polyp to which the anchor belongs to. The ground-truth distribution is defined as:

$$\boldsymbol{p}_i = \frac{\mathbb{1}_{\text{match}}(\boldsymbol{a}, \boldsymbol{b}_i) \cdot \exp(-\lambda \boldsymbol{d}_i)}{\sum_{j=1}^{K} \mathbb{1}_{\text{match}}(\boldsymbol{a}, \boldsymbol{b}_j) \cdot \exp(-\lambda \boldsymbol{d}_j)} \tag{2}$$

where $\mathbb{1}_{\text{match}}(\cdot, \cdot)$ is an indicator function identifying matches and $\lambda$ is a scaling factor that controls the influence of temporal distance. This formulation ensures that matching candidates closer in time to $\boldsymbol{a}$ receive higher weights independently of their visual appearance, while more distant matches are exponentially suppressed. We provide a visualization in Fig. 2 (right). The loss is then given by the cross-entropy between $\mathbf{p}$ and $\mathbf{q}$:

$$\mathcal{L} = H(\mathbf{p}, \mathbf{q}) = -\sum_{i=1}^{K} p_i \log q_i. \tag{3}$$

### 2.2 Polyp counting via clustering with temporal penalties

Similarly to [24], this module aims to cluster together tracklets that correspond to the same polyp entity. Alongside to visual similarity, we introduce another measure of similarity based on temporal information. Specifically, we integrate

a temporal penalty term that penalizes associations between tracklets that are far apart in time, and thus are unlikely to belong to the same cluster. The Fig. 1 (Bottom) depicts this component in details.

Formally, given a set of tracklet $\{\boldsymbol{t}_i\}$ from a video with $N$ tracklets, we compute the embedding $\boldsymbol{e}_i$ for each of them through the visual-temporal encoder. Then, we obtain $\boldsymbol{p}_i$, the vector that encodes the position of $\boldsymbol{t}_i$ against all $\boldsymbol{t}_j$ in video, normalized by video length. In order to perform the clustering, we obtain two matrices. The first is a visual similarity score matrix $V \in \mathbb{R}^{N \times N}$, that models appearance-based similarity. $V$ is obtained by computing all pairwise similarity scores $V_{i,j} = \text{sim}(\boldsymbol{e}_i, \boldsymbol{e}_j)$, where sim is a similarity function, e.g. cosine similarity. The second is a temporal adjacency matrix $T \in \mathbb{R}^{N \times N}$, $T_{i,j} = e^{-\gamma|\boldsymbol{p}_i - \boldsymbol{p}_j|}$ where $\gamma$ is the penalty factor. We normalize both $V$ and $T$ to get values in $[0, 1]$.

Finally, we compute $S = V \oplus T$ by combining $V$ and $T$, e.g. by summing the contributions: $V \oplus T = \alpha \cdot V + (1 - \alpha) \cdot T$ and clustering can be performed. Literature associates tracklets together either by using a threshold based approach [15] or unsupervised clustering algorithms [24]. Given the superior performance shown by the latter method, especially by Affinity Propagation [9], we adopt this algorithm for our approach. From clusters, polyp counting can be finally derived.

## 3 Experiments

### 3.1 Datasets and evaluation protocol

We construct our training set from publicly available datasets, and follow the evaluation framework proposed in [24]. Specifically, we extend the training set by combining 85 polyps from REAL-Colon, already used in [24], with 160 from LDPolyp [20], 100 from SUN [21] and 39 from PolypSet [18], for a total of 384 polyps. Each dataset provides polyp detections, but (polyp, video) annotations are available only in REAL-Colon. For this reason, we cannot extend the evaluation set to other datasets. In [24], authors propose a validation set with 10 videos (23 polyps) and a test set with 9 videos (24 polyps). Although REAL-Colon is a multi-centric dataset, the limited size of the validation set may lead to unreliable estimates of test performance. To address this, we use a leave-one-out cross-validation (LOOCV) strategy, ensuring a more reliable and robust assessment. Following [15,24], we measure the fragmentation rate (FR), defined as the average number of tracklets polyps are split into. The FR is a number $\geq 1$, where lower values mean lower fragmentation. Formally, we define the FR as a function of $T$ a set of tracklets and $E$ a set of polyp entities. Given a set of clustered tracklets $C$, with $|C| \leq |T|$, we can compute the fragmentation rate as $FR = |C|/|E|$. A perfect re-association would have $|C| = |E|$, thus $FR = 1$. To account for false positive associations, we select the clustering hyper-parameters that yield the lowest fragmentation rate (FR) at fixed false positive rate (FPR) $\rho = 0.05$ on the combined validation set of 18 videos and evaluate on the remaining one, reporting the average FR and FPR across all 19 videos. Thus, $FR = |\hat{C}|/|E|$, where $\hat{C}$ is the set of clustered tracklets that minimizes $|FPR(C) - \rho|$.

| Method | FR ↓ | FPR ↓ |
|---|---|---|
| No clustering | 20.74 ±20.78 | 0.0000 ±0.0000 |
| SFE [15] | 9.37 ±6.28 | 0.0314 ±0.0618 |
| MVE [15] | 8.42 ±4.50 | 0.0449 ±0.1361 |
| TPC [24] | 5.10 ±3.43 | 0.0249 ±0.0408 |
| **Ours** | **2.32 ±1.80** | **0.0225 ±0.0420** |

Table 1: Mean and standard deviation results of the Fragmentation Rate (FR) and False Positive Rate (FPR) on REAL-Colon using LOOCV. We report results for models trained on the combination of public datasets composed by REAL-Colon, LDPolyp, SUN and PolypSet. The "No clustering" baseline reflects the initial fragmentation rate before any clustering.

### 3.2 Implementation details

The visual encoder is implemented following [15]. For comparability with previous works [15,24], we use ResNet-50 as frame encoder pre-trained on ImageNet [7], and set the model size $d = 128$. Due to memory limitations, we set fragment length $\kappa = 8$ and batch size to 56. Number of different polyps in a batch is set to 3, i.e. 14 views per polyp (unless otherwise specified). We set $\lambda = 1$. We select the bounding box scaling factor $\psi$ in $\{1, 2, 3, 5, 10\}$ and chose 5. For fair comparison, all methods were trained and evaluated under the same setting. We construct tracklets from ground-truth polyp detections to avoid tracker noise and enable fair comparison with prior work. Each tracklet is a sequence of consecutive frames from the same polyp entity, with an Intersection Over Union (IoU) of at least 0.1 between consecutive frames. To avoid redundancy from high fps videos in tracklets, we encode one frame every four. We train all models on the combined dataset from public sources. Our runtime machine is a node with 6 cores, 60GB of RAM and one NVIDIA RTX A6000 GPU with 48GB of VRAM. We select clustering hyper-parameters with a grid-search. For the threshold algorithm we vary the threshold in $[0, 0.01, \ldots, 1]$. For Affinity Propagation we select the preference parameter in $[-5, -4.75, \ldots, 5]$, while for our temporal clustering we select $\gamma \in [0.1, 0.2, \ldots, 0.9] \cup [1, 1.375, \ldots, 10]$ and $\alpha \in [0, 0.05, \ldots, 1]$ along with Affinity Propagation's preference.

### 3.3 Results

We compare our approach to the state-of-the-art [15,24] and present results in Tab. 1. For fair comparison, we train previous methods on the same combined dataset used in our experiments. Our method outperforms all existing approaches, achieving 3.6 and 4.0 times lower fragmentation rate with respect to SFE and MVE from [15]. Notably, TPC [24] also employs Affinity Propagation for clustering, the same underlying algorithm used in our approach, but our method yields a 2.2x improvement thanks to the combination of our novel visual-temporal representation learning and temporal clustering.

| N. views per polyp | N. polyps in batch | FR ↓ | FPR ↓ |
|---|---|---|---|
| 2 | 28 | 2.57 ±2.07 | 0.0276 ±0.0576 |
| 4 | 14 | 3.97 ±3.13 | 0.0417 ±0.0987 |
| 7 | 8 | 2.36 ±1.77 | 0.0325 ±0.0642 |
| **14** | **3** | **2.32 ±1.80** | **0.0225 ±0.0420** |
| 28 | 2 | 4.54 ±3.86 | 0.0317 ±0.0557 |

Table 2: Fragmentation rate results on REAL-Colon by number of views per polyp in a batch of 56 elements.

### 3.4 Ablation study

*Effect of multiple positive views per polyp on representation learning.* Previous works [15,24] follows the traditional contrastive learning framework [6], adopting two positive views per polyp. Instead, our supervised setting allows to derive multiple meaningful views from polyp entities. We evaluate the impact of this parameter on the quality of representations. With a fixed batch capacity of 56 elements, we investigate the number of views per polyp $K$ in $\{2, 4, 7, 14, 28\}$, corresponding to $56/K = 28$, 12, 8, 3, and 2 polyps per batch, respectively. Our results show that increasing the number of views helps reducing the fragmentation rate (FR) and false positive rate (FPR), with the lowest FR (2.32) and FPR (0.0225) achieved with 14 views (3 polyps in batch).

*Impact of temporal information on polyp counting.* We investigate the impact of temporal information on both tracklet representation learning and clustering in Tab. 3. On the top, we can note the superior performance of our temporally-aware supervised loss over the standard supervised loss, which in turn improves on the self-supervised loss. Since the clustering algorithm remains unchanged, the observed improvement can be attributed to a more discriminative embedding space. On the bottom, the results show that incorporating temporal information into the clustering process (Temporal-AP) significantly reduces fragmentation rate (FR) suggesting that the temporal association penalty helps form more accurate clusters. Notably, both visual and temporal cues contribute to the final clustering. The weighting factor $\alpha$, which balances the influence of the two modalities, averages $0.4833 \pm 0.1999$ across the LOOCV folds, indicating that their combination is essential for optimal performance.

## 4 Conclusion

In conclusion, we propose a novel polyp counting approach using supervised contrastive learning to enhance tracklet representation. By incorporating temporally-aware soft targets, our method improves intra-polyp invariance and inter-polyp separability. Temporal constraints in the clustering stage further refine tracklet associations, enhancing reliability. Extensive evaluation on public datasets, including an expanded training set and leave-one-out cross-validation, shows that

| Loss | Clustering | FR ↓ | FPR ↓ |
|---|---|---|---|
| Self-supervised | Temporal-AP | 10.47 ±19.96 | 0.0210 ±0.0248 |
| Supervised | Temporal-AP | 3.96 ±2.68 | 0.0252 ±0.0507 |
| **Temporally-aware** | **Temporal-AP** | **2.32 ±1.80** | **0.0225 ±0.0420** |
| Temporally-aware | AP | 4.18 ±2.86 | 0.0259 ±0.0397 |
| **Temporally-aware** | **Temporal-AP** | **2.32 ±1.80** | **0.0225 ±0.0420** |

Table 3: Ablation results on REAL-Colon by loss functions and clustering algorithms. AP=Affinity Propagation, Temporal-AP=AP with temporal penalty.

our method outperforms previous approaches, reducing fragmentation by a factor of 2.2.

**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. and A.C. are affiliated with Cosmo Intelligent Medical Devices, the developer of the GI Genius medical device.

## References

1. Amano, T., Nishida, T., Shimakoshi, H., Shimoda, A., Osugi, N., Sugimoto, A., Takahashi, K., Mukai, K., Nakamatsu, D., Matsubara, T., et al.: Number of polyps detected is a useful indicator of quality of clinical colonoscopy. Endoscopy International Open **6**(07), E878–E884 (2018)
2. 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)
3. Biffi, C., Salvagnini, P., Dinh, N.N., Hassan, C., Sharma, P., Cherubini, A.: A novel ai device for real-time optical characterization of colorectal polyps. NPJ digital medicine **5**(1), 84 (2022)
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, Q., Cai, S., Cai, C., Yu, Z., Qian, D., Xiang, S.: Colo-scrl: Self-supervised contrastive representation learning for colonoscopic video retrieval. In: Proceedings of the IEEE International Conference on Multimedia and Expo. pp. 1056–1061 (2023)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A Simple Framework for Contrastive Learning of Visual Representations. In: Proceedings of the International Conference on Machine Learning. pp. 1597–1607 (2020)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)

8. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceeding of the Association for Computational Linguistics. pp. 4171–4186 (2019)
9. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. science **315**(5814), 972–976 (2007)
10. Gimeno-García, A.Z., Hernández-Pérez, A., Nicolás-Pérez, D., Hernández-Guerra, M.: Artificial intelligence applied to colonoscopy: Is it time to take a step forward? Cancers **15**(8), 2193 (2023)
11. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)
12. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)
13. 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. 9726–9735 (2020)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 770–778 (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. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)
17. Lee, S., Park, T., Lee, K.: Soft contrastive learning for time series. In: Proceedings of the International Conference on Learning Representations (2024)
18. 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)
19. Lux, T.J., Saßmannshausen, Z., Kafetzis, I., Sodmann, P., Herold, K., Sudarevic, B., Schmitz, R., Zoller, W.G., Meining, A., Hann, A.: Assisted documentation as a new focus for artificial intelligence in endoscopy: the precedent of reliable withdrawal time and image reporting. Endoscopy **55**(12), 1118–1123 (2023)
20. Ma, Y., Chen, X., Cheng, K., Li, Y., Sun, B.: Ldpolypvideo benchmark: a large-scale colonoscopy video dataset of diverse polyps. In: Medical Image Computing and Computer Assisted Intervention. pp. 387–396 (2021)
21. Misawa, M., Kudo, S.e., 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 (with video). Gastrointestinal endoscopy **93**(4), 960–967 (2021)
22. Nie, M.Y., An, X.W., Xing, Y.C., Wang, Z., Wang, Y.Q., Lü, J.Q.: Artificial intelligence algorithms for real-time detection of colorectal polyps during colonoscopy: a review. American Journal of Cancer Research **14**(11), 5456 (2024)
23. Nogueira-Rodríguez, A., Domínguez-Carbajales, R., López-Fernández, H., Iglesias, Á., 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. Parolari, L., Cherubini, A., Ballan, L., Biffi, C.: Towards polyp counting in full-procedure colonoscopy videos. In: Proceedings of the IEEE International Symposium on Biomedical Imaging. pp. 1–5 (2025)
25. Qian, R., et al.: Spatiotemporal contrastive video representation learning. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 6964–6974 (2021)
26. Sobal, V., Ibrahim, M., Balestriero, R., Cabannes, V., Bouchacourt, D., Astolfi, P., Cho, K., LeCun, Y.: X-sample contrastive loss: Improving contrastive learning with sample similarity graphs. In: Proceedings of the International Conference on Learning Representations (2025)
27. Tavanapong, W., Oh, J., Riegler, M.A., Khaleel, M., Mittal, B., De Groen, P.C.: Artificial intelligence for colonoscopy: Past, present, and future. IEEE journal of biomedical and health informatics **26**(8), 3950–3965 (2022)
28. Tian, Y., Fan, L., Isola, P., Chang, H., Krishnan, D.: Stablerep: Synthetic images from text-to-image models make strong visual representation learners. Advances in Neural Information Processing Systems **36**, 48382–48402 (2023)
29. 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)
30. Xiang, S., Liu, C., Ruan, J., Cai, S., Du, S., Qian, D.: Vt-reid: Learning discriminative visual-text representation for polyp re-identification. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 3170–3174. IEEE (2024)