Semi-supervised Quality Evaluation of Colonoscopy Procedures

Idan Kligvasser George Leifman Roman Goldenberg Ehud Rivlin Michael Elad
Verily

Abstract

Colonoscopy is the standard of care technique for detecting and removing polyps for the prevention of colorectal cancer. Nevertheless, gastroenterologists (GI) routinely miss approximately 25% of polyps during colonoscopies. These misses are highly operator dependent, influenced by the physician skills, experience, vigilance, and fatigue. Standard quality metrics, such as Withdrawal Time or Cecal Intubation Rate, have been shown to be well correlated with Adenoma Detection Rate (ADR). However, those metrics are limited in their ability to assess the quality of a specific procedure, and they do not address quality aspects related to the style or technique of the examination. In this work we design novel online and offline quality metrics, based on visual appearance quality criteria learned by an ML model in an unsupervised way. Furthermore, we evaluate the likelihood of detecting an existing polyp as a function of procedure quality and use it to demonstrate high correlation of the proposed metric to polyp detection sensitivity. The proposed online quality metric can be used to provide real time quality feedback to the performing GI. By integrating the local metric over the withdrawal phase, we build a global, offline quality metric, which is shown to be highly correlated to the standard Polyp Per Colonoscopy (PPC) quality metric.

1. Introduction

Screening for colorectal cancer is highly effective, as early detection is within reach, making this disease one of the most preventable. Today's standard of care screening method is optical colonoscopy, which searches the colon for mucosal abnormalities, such as polyps. However, performing a thorough examination of the entire colon surface using optical colonoscopy is challenging, which may lead to a lower polyp detection rate. Recent studies have shown that approximately 25% of polyps are routinely missed during colonoscopies [1].

The success (diagnostic accuracy) of a colonoscopy procedure is highly operator dependent. It varies based on the performing physician skills, experience, vigilance, fatigue, and more. To ensure high procedure quality, various quality metrics are measured and monitored. E.g., the Withdrawal Time (time from the colonoscope reaching cecum to removal of the instrument from the patient) metric was shown to be highly correlated to Adenoma Detection Rate (ADR) [8, 17, 20, 21, 22, 24]. Another quality metric – Cecal Intubation Rate (proportion of colonoscopies in which the cecum is intubated) – is considered important to ensure good colon coverage.

Most of these existing metrics are relatively easy to compute, but can provide only limited data on the quality of a specific procedure, and are typically used aggregatively for multiple sessions. Some studies [18] suggest that there are other factors that impact the polyp detection rate. For example, one may wish to distinguish between a good and bad colonoscope motion patterns, or assess the style of the examination. The hypothesis is that a better inspection style yields more informative visual input, which results in a better diagnostic accuracy.

In this work we propose a novel quantitative quality metric for colonoscopy, based on the automatic analysis of the induced video feed. This metric is computed locally in time, measuring how informative and helpful for colon inspection a local video segment is. As this instantaneous quality is very subjective and difficult to formulate, human annotation is problematic and ill-defined. Instead, we let an ML model build a meaningful visual data representation in a fully unsupervised way, and use it to construct a metric highly correlated with the clinical outcome. First, we learn visual representations of colonoscopy video frames using contrastive self-supervised learning. Then, we perform cluster analysis on these representations and build a linear classifier base on these cluster assignments, bearing a strong correlation with polyp detection, which can serve as an indicator for "good-quality" video segments.

While the proposed approach resembles the one proposed in [9], the addressed problems are markedly different, as [9] does phase detection in colonoscopy. There are other works aiming to learn frame representations in colonoscopy videos. However, those descriptors are usually associated with polyps, and used for polyp related tasks - tracking, re-identification [4, 25], optical biopsy [23], etc.

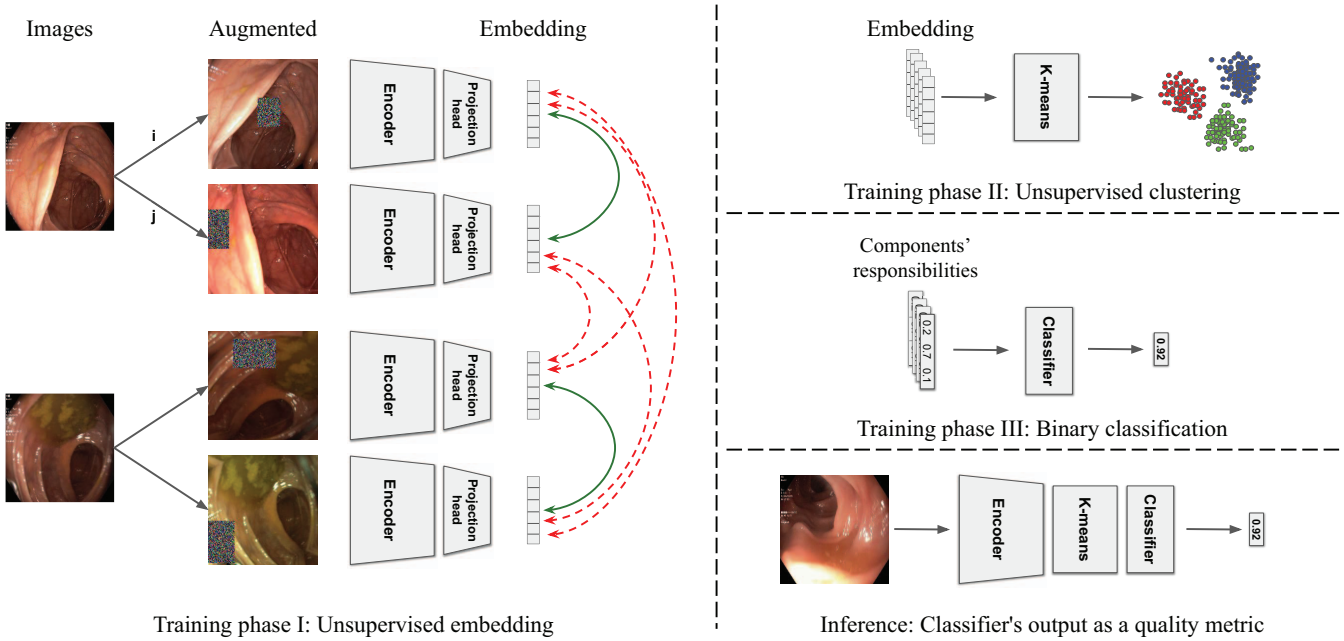

Figure 1: **Method overview.** *(Left)* Two augmented views for each frame are used to train the encoder and the projection head using contrastive learning. *(Right top)* Feature representations are directly clustered into semantically meaningful groups using K-means. *(Right middle)* Learning clusters' associations. *(Right bottom)* At inference time, cluster attributes are leveraged for quality metric evaluation.

By measuring the duration of good quality video segments over the withdrawal phase of the procedure, we derive a new offline colonoscopy quality metric. We show that this measure is strongly correlated to the Polyps Per Colonoscopy (PPC) quality metric. Moreover, we show how the real-time measurement of the quality of a colonoscopy procedure can be used to evaluate the likelihood of detecting a polyp at any specific point in time during the procedure.

The rest of the paper is structured as follows: Section 2 introduces the encoding and clustering of frames, which are then utilized to define quality metrics. Section 3 showcases the evaluation of the proposed metrics, along with ablation studies. Lastly, Section 4 summarizes the findings and discusses potential future directions.

2. Method

Our goal is to learn a colonoscopy quality metric through the identification of temporal intervals in which effective polyp detection is possible. We start by learning the colonoscopy video frame embedding using self-supervised learning, followed by a cluster analysis. Using those clusters, we learn a "good" frame classifier, which then serves as the basis for both global (offline) and local (online) quality metrics. The end-to-end framework is described in the following sections, and illustrated in Figure 1.

2.1. Frame Encoding

We start from learning visual representations of colonoscopy frames using contrastive learning. We use SimCLR [5], which maximizes the agreement between representations of two randomly augmented versions of the same frame, while pushing away the representations of other frames (see Fig. 1). Specifically, frame x_i is randomly augmented, resulting in two correlated views, x_i^1 and x_i^2, considered as a positive pair. These views are fed to an encoder $f_\theta(\cdot)$ and projection layer $g_\phi(\cdot)$, yielding the embedding vector $z_i^a = g_\phi(f_\theta(x_i^a))$ $(a = 1, 2)$. Given a batch of N frames, the contrastive loss referring to the i-th frame is given by:

$$\ell(z_i^1, z_i^2) = -log \frac{exp(sim(z_i^1, z_i^2)/\tau)}{\sum_{k \neq i} \sum_{a=1}^{2} \sum_{b=1}^{2} exp(sim(z_i^a, z_k^b)/\tau)}, \quad (1)$$

where τ is a temperature parameter, and sim is the cosine similarity defined as $sim(u, v) = u^T v / \|u\| \|v\|$. We use ResNet-RS50 [3] for the encoder and a simple MLP with one hidden layer for the projection layer, as suggested in [5].

Our training data consists of $1M$ frames randomly sam-

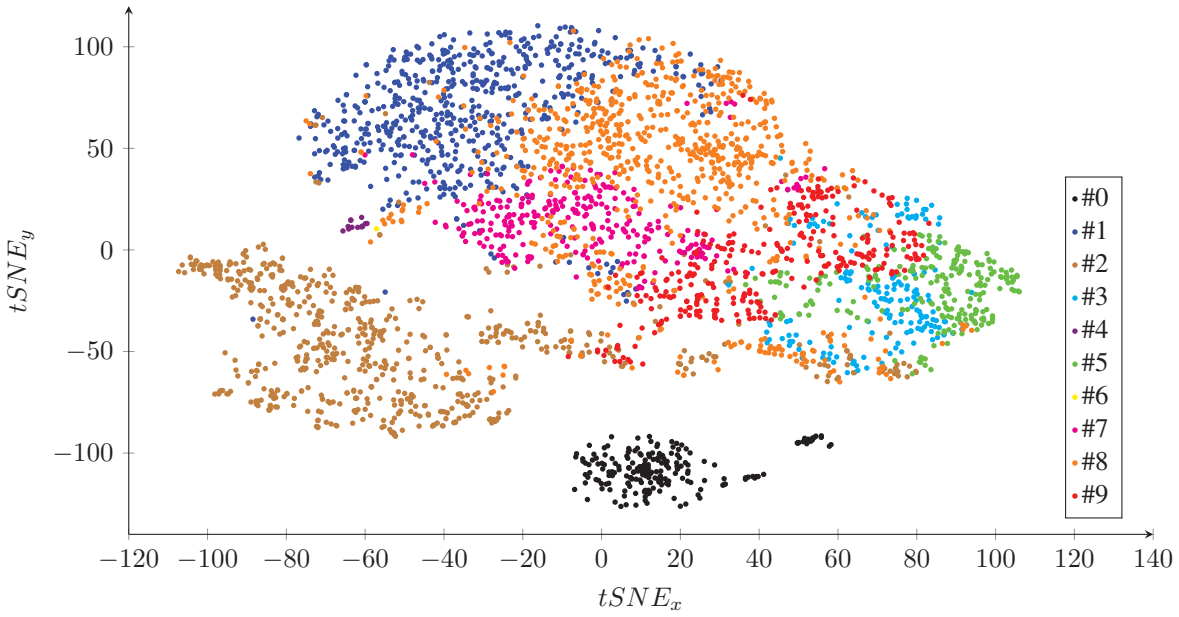

Figure 2: **T-SNE plot of frame embeddings.** K-means clusters are color coded.

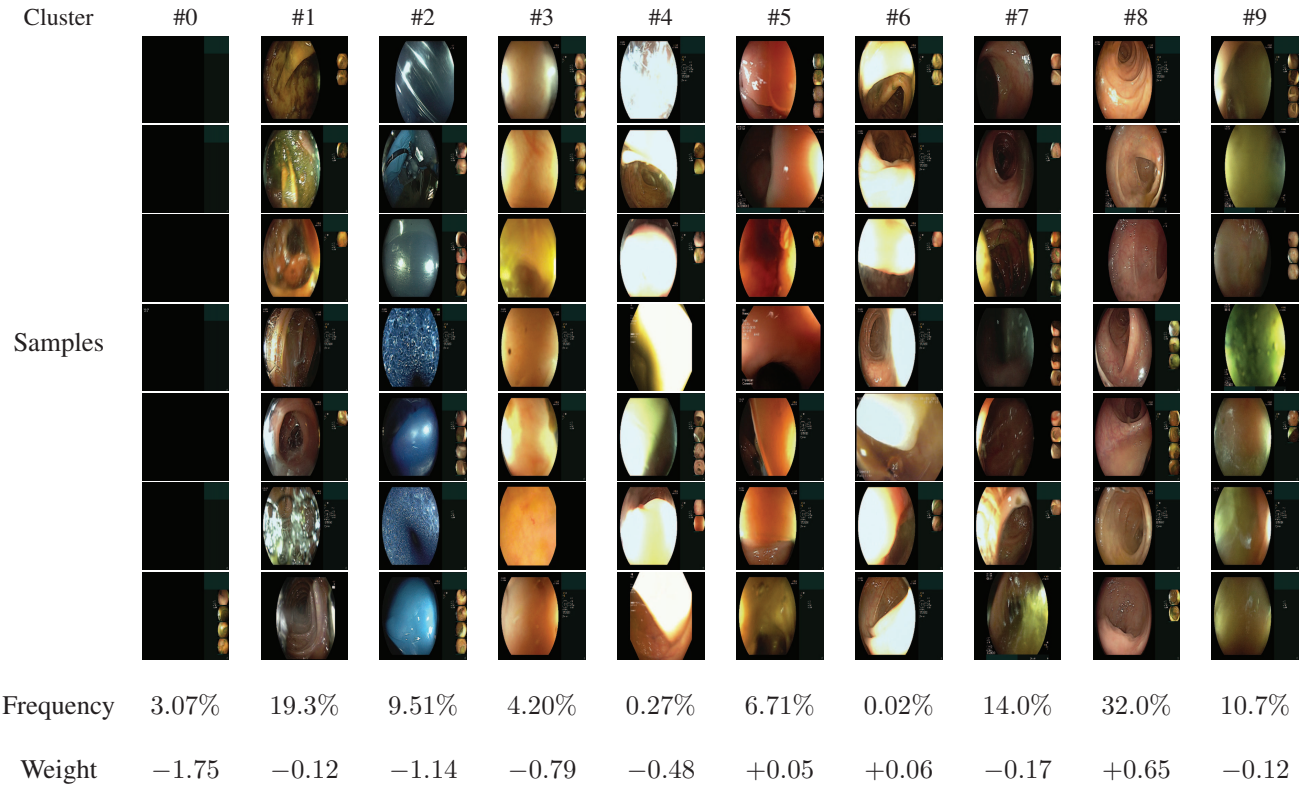

Figure 3: **Clusters visualization.** Random selection of frames from each cluster along with their corresponding relative frequency. Cluster #0 refers to out-of-body content, thus given a negative weight. Conversely, cluster #8 comprises tubular clean frames and thus associated with polyp detection with the highest positive weight.

pled from 2243 colonoscopy videos. Since the designed metric is supposed to be used for predicting the chance of detecting a polyp, it is not expected to be used on frames where the polyp is detected or treated. Therefore, we exclude such frames from the training set, by detecting them automatically using off-the-shelf polyp and surgical tool detectors [11, 12, 13]. As the negative pairs (red arrows in Fig. 1/Left) are utilized within the same batch as the positive ones (green arrows), in each batch, each sample includes frames from its own temporal neighborhood for a more challenging negative mining.

For augmentation we use standard geometric transformations (resize, rotation, translation), color jitter, and the Cutout [6] with the Gaussian noise filling. The augmented views are illustrated in Fig. 1/Left.

2.2. Frame Clustering

The second step in our scheme is clustering the learned representations $f_\theta(x_i)$ into K(=10) clusters using k-means [14]. While the standard k-means does a hard assignment of each frame to its corresponding cluster, we use a soft alternative based on the distance between the frame descriptor to cluster centers. Namely, we define the probability of the i-th frame to belong to the k-th cluster by

$$r_{i,k} = Prob(f_\theta(x_i) \in k) \sim \left[\frac{1}{\|f_\theta(x_i) - c_k\|_2^2}\right]^\alpha \quad \text{for } k = 1, 2, \ldots, K, \quad (2)$$

where $\{c_k\}_{k=1}^K$ are the cluster centers, $\alpha = 16$, and vector $r_i = (r_{i,k})_{k=1}^K$ is normalized to sum 1. Figure 2 shows the t-SNE projection of frame embeddings with k-means clusters color coded. Interestingly, the samples are clustered into relatively compact, meaningful groups. Figure 3 presents a random selection of frames from each cluster. One can see that clusters $1, 2$ and 7 contains inside-body informative frames. In contrast, clusters $0, 3, 4, 5, 6, 8$ and 9 contain non-informative outside-body and inside-body frames. Please see the Supplemental Materials for more visual examples.

2.3. Online (Local) Quality Metric

Based on the learned frame embeddings and clusters, we now design an online (local) quality metric.

As our objective is to link the visual appearance to polyp detection, we learn a metric that predicts one from the other. Namely, we learn a function $Q(\cdot)$ that maps frame x_i appearance encoded by the vector r_i (see Eq. 2) to the chance of detecting a polyp in the following frames.

More precisely, we average the r_i over a video segment of 10 sec to get $\overline{r_i}$, and train a binary classifier $Q(\overline{r_i})$ to predict the detection of a polyp in the following 2 seconds.

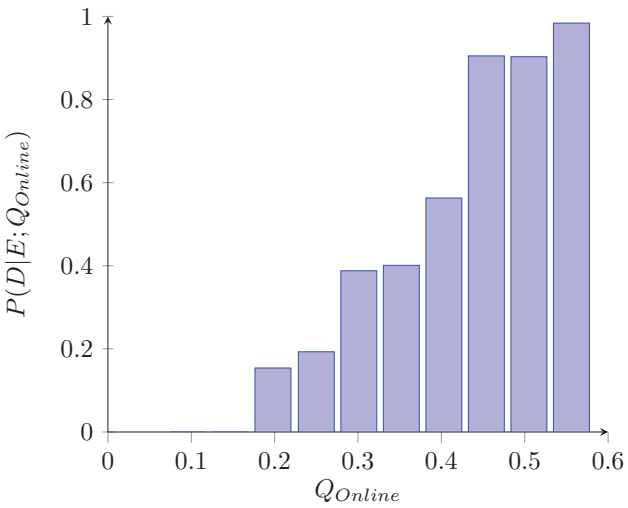

Figure 4: The likelihood of detecting an existing polyp in a short video segment as a function of local quality metric Q. As one can see, the proposed quality metric Q correlates very well with the polyp detection sensitivity (PDS).

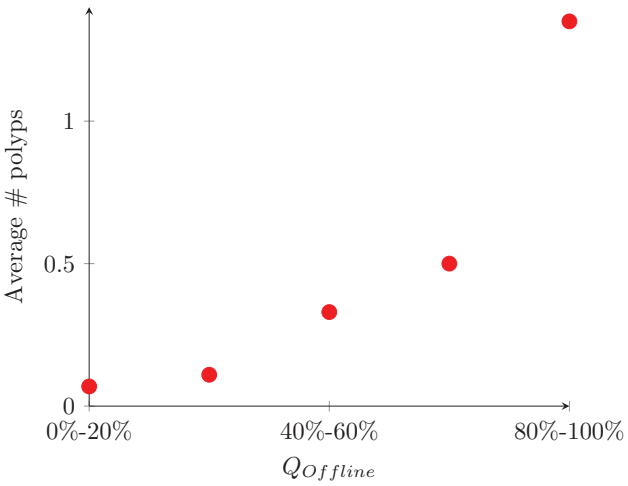

Figure 5: **Q_{Offline} during the withdrawal phase.** The relationship between the proposed offline quality measure and the actual number of polyps detected, when Q_{Offline} observations are divided into five equal-sized groups. One can observe a strong correlation between the Q_{Offline} the PPC metric.

The training set for the classifier is built from a set of 2243 colonoscopy videos annotated for the location of polyps. 1086 intervals of 10 seconds before the appearance of polyps are sampled from the training set as positive samples, and another 1086 random intervals sampled as negative samples. The $Q(\cdot)$ is implemented as a binary

classifier with a single linear layer and trained with Adam optimizer [10] for 500 epochs, using a batch size of 64.

By examining the weights of the linear classifier, we can gain some insights into how it estimates the likelihood of detecting a polyp within the next 2 seconds. Each cluster is assigned a weight as shown in Figure 3. E.g., class #0 refers to out-of-body content, thus given a negative weight. Conversely, class #8 refers to tubular, clean frames and, thus, is given the highest positive weight for polyp detection.

While the $Q(\cdot)$ achieves only a mediocre classification accuracy (i.e. polyp detection prediction) accuracy of 64% on the test-set (indeed, it is very difficult to predict a detection of a polyp when it is not known that the polyp is there), we will show in the following sections that it can still be used as a quality metric.

2.4. From Quality Metric to the Chance to Detect a Polyp

We would like to assess the chance of detecting a polyp (if it exists) at time t as a function of the procedure quality Q in the preceding time interval $[t-\Delta t, t]$. Let us denote the event of having a polyp in the colon at time t as E ("exists"), and the event of detecting it as D ("detected"). For this analysis we will treat the quality metric Q from the previous section, as a random variable in the range $[0, 1]$ measuring the quality of the procedure in the time interval $[t - \Delta t, t]$.

We are interested to estimate the following probability:

$$P(D|E,Q) = \frac{P(E,Q|D)P(D)}{P(E,Q)} = \\ = \frac{P(Q|D)P(E|Q,D)P(D)}{P(E,Q)}, \quad (3)$$

representing the chance of detecting a polyp if it exists as a function of quality. In the above, the first equality uses the Bayes rule, and the second exploits the chain probability relationship. We know that physicians rarely mistake a non-polyp for a polyp, implying that $P(E|Q,D) \approx 1$. Then, assuming the independence between the existence of the polyp (E) and the quality of the procedure (Q), Eq. 3 becomes

$$P(D|E,Q) \approx \frac{P(Q|D)P(D)}{P(Q)P(E)} \quad (4)$$

As mentioned above, the incidence of polyp detection false alarms in colonoscopy is negligible, hence the ratio $P(D)/P(E)$ can be interpreted as the average polyp detection rate/sensitivity (PDS). From the literature, we know that polyp miss-rate in colonoscopy is about $20 - 25\%$ [1]. Hence, $P(D)/P(E)$ can be approximated as $0.75 - 0.8$, regardless of Q.

Therefore, to compute $P(D|E,Q)$, all we need to do is approximate $P(Q)$ and $P(Q|D)$. This can be done empirically by estimating the distribution of Q in random intervals and in intervals preceding polyps for $P(Q|D)$.

2.5. Offline Quality Metric (Post-Procedure)

We would like to design an offline quality indicator based on the above online measure Q. We define the following quality metric by integrating Q over the entire withdrawal phase,

$$Q_{\text{Offline}} = \sum_{i \in \text{withdrawal}} Q\left(r_i\right). \quad (5)$$

3. Experiments

3.1. Online Quality Metric Evaluation

We would like to evaluate how relevant the proposed online quality metric Q is to the ability of detecting polyps. We do that by estimating the likelihood of detecting an existing polyp $P(D|E,Q)$ as a function of Q. The higher the correlation between Q and $P(D|E,Q)$, the better Q is as a local colonoscopy quality metric.

As discussed above $P(D|E,Q) \propto P(Q|D)/P(Q)$. Both $P(Q|D)$ and $P(Q)$ can be estimated empirically: For $P(Q)$ we build a 10-bin histogram of Q measured in 543 randomly chosen colonoscopy video segments $10sec$ long. The same is done for $P(Q|D)$, but with 543 video segments preceding a polyp.

The estimated $P(D|E,Q)$ is depicted in Figure 4. As one can see, the proposed quality metric Q correlates very well with the polyp detection sensitivity (PDS). Q can be computed online and provided as a real time feedback to the physician during the procedure.

3.2. Offline Quality Metric Evaluation

We would like to evaluate the effectiveness of the proposed offline quality metric Q_{Offline} in predicting the polyp detection sensitivity.

To do so, we compute Q_{Offline} for 500 annotated test set colonoscopies. We sort the cases in the increasing order of Q_{Offline}, and split them into 5 bins - 100 cases each, from lower Q_{Offline} to higher. For each bin we compute the average Polyps Per Colonoscopy (PPC) metric. The resulting histogram is shown in Figure 5. One can observe a strong correlation between the Q_{Offline} and the PPC metric.

Figure 6 shows the distribution of procedures with (red) and without detected polyps (blue), as the function of Q_{Offline}. One can see that higher Q_{Offline} are more likely to correspond to procedures with detected polyps.

The evaluations above suggest that the proposed quality metric Q_{Offline} is highly correlated to polyp detection sensitivity (PPS). It is important to note that high Q_{Offline} for any specific procedure does not mean that there is a high chance of finding a polyp in that procedure, as we don't

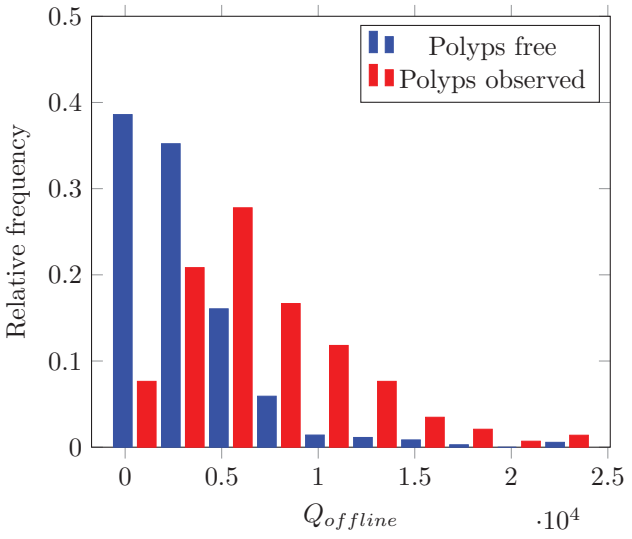

Figure 6: **$Q_{Offline}$ during the withdrawal phase.** Procedures with high $Q_{Offline}$ values are likely to have polyps. One can see that higher $Q_{Offline}$ are more likely to correspond to procedures with detected polyps.

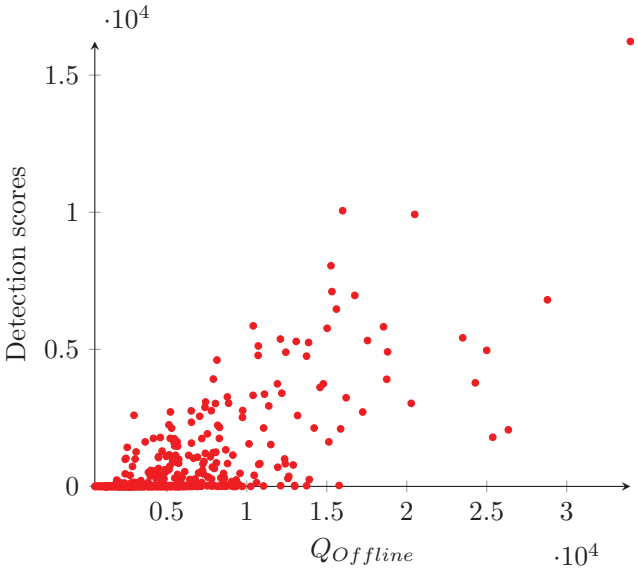

Figure 7: **$Q_{Offline}$ during the withdrawal phase.** Analyzing the relationship between the suggested metric and the absolute polyp frame detection scores. As can be seen, the proposed measure is strongly correlated with the frame detection scores.

know if there are any polyps there and how many. What it does mean, is that if there is a polyp, there is a high chance it will be detected.

In the context of individual procedures and the proposed

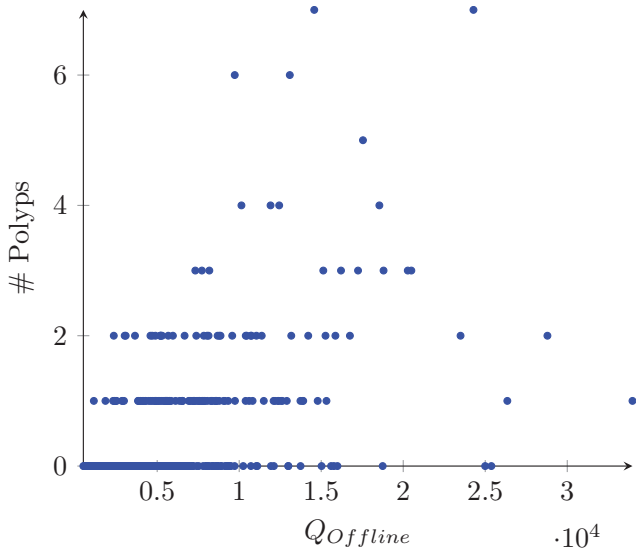

Figure 8: **$Q_{Offline}$ during the withdrawal phase.** Examining the association between the proposed metric and annotated number of polyps found in each procedure in each procedure. As can be seen, our findings reveal a robust association between the proposed measure and the presence of polyps.

metric relationships, our evaluation focuses on measuring the absolute polyp frame detection scores and the annotated number of polyps detected in each of the 500 colonoscopies within the test set. We examine the relationship between the absolute polyp frame detection score and the suggested metric in Fig 7, while Fig 8 illustrates the connection between the suggested metric and the annotated number of polyps found in each procedure. Notably, our findings reveal a robust association between the proposed measure and the presence of polyps, even within a single procedure. It is important to note that a low value of this quality measure does not necessarily indicate poor performance if a procedure does not reveal any polyps. Instead, it could signify a healthy patient (has no polyps) and a well-conducted inspection.

3.3. Experimental Setups

This section elucidates the technical intricacies associated with our approach. The details encompass various aspects, including data collection and the implementation of algorithms.

Dataset: Our dataset comprises a total of 2743 full colonoscopy videos obtained from five distinct hospitals in the United States, Japan and Israel. Each video is captured in Full HD resolution (1080p) with frame rates ranging from 25 to 60 frames per second (FPS). The duration of each procedure varies, with some lasting only a few minutes and the

General information		
	# of hospitals	5
	# of procedures	2743
Gender		
	Female	54%
	Male	46%
Age groups		
	25-40	12%
	41-55	26%
	56-70	48%
	70+	14%
Annotations		
	Polyp intervals	1586

Table 1: **Dataset summary.** Our dataset consists of full 2743 colonoscopy videos taken from 5 hospitals in the United States, Japan and Israel.

longest extending up to an hour and a half. In compliance with privacy regulations, we removed any identifiable information, including operator IDs, to safeguard confidentiality. Other demographic parameters are summarized in Table 1.

Implementation Details: The proposed model was implemented using the Tensorflow framework, utilizing multiple Tensor Processing Units (TPUs) for efficient training. For the training of frame encoding (Section 2.1), we adopted the SIMCLR [5] training scheme as the foundation. The Adam optimizer [10], with its default settings, was employed for training the network. The initial learning rate was set to 10^{-3} and progressively decayed to 10^{-5} during the training process. Throughout the training, we maintained a fixed mini-batch size of 512 and conducted $100k$ optimization steps. Subsequently, for training the linear classifier 2.3, we once again employed the Adam optimizer with an initial learning rate of $10^{-3}3$, which was gradually reduced to 10^{-5}. A mini-batch size of 64 was utilized, and the optimization process spanned $16k$ steps. The specific configurations and parameters were chosen to optimize the performance and convergence of the proposed model.

3.4. Clustering Ablation Study

Within this section, our focus is on investigating the selection of the clustering algorithm and its corresponding hyperparameters in the frame clustering stage of our method. To conduct this analysis, we randomly sampled $10k$ training representations and $5k$ test representations. Initially, we choose the clustering algorithm based on the entropy score, which allows us to determine the algorithm that best suits our needs. Subsequently, we determine the optimal number of clusters using the Bayesian information criterion (BIC) [19] measure. To ensure the reliability of our findings, we repeat the evaluation process 10 times for each configuration and calculate the mean and standard deviation of the results. This approach enables us to obtain robust and informative outcomes.

Algorithm	Components	Entropy
DBSCAN	–	0.39 ± 0.13
OPTICS	–	0.77 ± 0.15
GMM	10	1.21 ± 0.21
BIRCH	10	1.41 ± 0.31
K-Means	5	1.24 ± 0.27
K-Means	**10**	$\mathbf{2.03 \pm 0.37}$

Table 2: **Algorithmic exploration.** The performance of popular clustering algorithms in terms of entropy scores and number of clusters. K-means outperforms other algorithms, producing a larger number of clusters, including many samples in each group.

In the initial stage of our ablation, we conducted a systematic evaluation of various clustering methods to identify the most suitable one. Our primary criterion for comparison is the level of disorder or entropy exhibited by each algorithm, as a higher entropy indicates a more diverse composition of clusters. The results of this evaluation are presented in Table 2, which provides a performance overview of popular clustering techniques such as DBSCAN [7], OPTICS [2], GMM [16], BIRCH [26], and K-Means [15]. As can be seen, K-means outperforms the other algorithms, producing a diverse number of groups.

In our study, the determination of the appropriate number of cluster components for the K-means algorithm is accomplished through the utilization of the Bayesian information criterion (BIC) [19], a statistical measure. Inspired by the principle of Occam's razor, which favors simpler explanations, the BIC incorporates a penalty for model complexity by introducing a term proportional to the number of parameters in the model. Consequently, when all other factors are held constant, a model with fewer parameters will exhibit a lower BIC compared to a model with a greater number of parameters. Our motivation in this regard stems from the necessity to establish an adequate number of clusters that accurately represent the colonoscopy domain. However, we also aim to constrain this quantity to encompass only generalized representations. This ensures that the clusters possess sufficient generality to be applicable across a wide range of procedures while still retaining the ability to discern pertinent information. The BIC scores corresponding to various choices of the number of components are presented in Fig 9. The results reveal that the optimal selection, based on the BIC criterion, occurs when $k = 9$ components are employed.

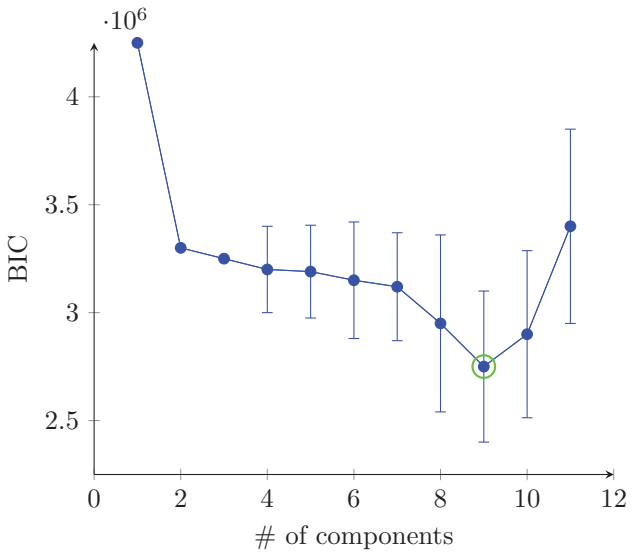

Figure 9: **Bayesian information criterion.** Bayesian information criterion (BIC) computed for 1 to 11 number of K-Means clusters. Results indicate that employing $k = 9$ components yields the optimal choice according to the Bayesian Information Criterion (BIC).

4. Conclusion

We proposed novel online and offline colonoscopy quality metrics, computed based on the visual appearance of frames in colonoscopy video. The quality criteria for the visual appearance were automatically learned by an ML model in an unsupervised way.

Using a Bayesian approach, we developed a technique for estimating the likelihood of detecting an existing polyp as a function of the proposed local quality metric. We used this likelihood estimation to demonstrate the correlation between the local quality metric and the polyp detection sensitivity. The proposed local metric can be computed online to provide a real time quality feedback to the performing physician.

Integrating the local metric over the withdrawal phase yields a global, offline quality metric. We show that the offline metric is highly correlated to the standard Polyps Per Colonoscopy (PPC) quality metric.

As the next step, we would like to estimate the impact of the proposed real time quality feedback on the quality of the procedure, e.g. by measuring its impact on the Adenoma Detection Rate (ADR) in a prospective study.

References

[1] Sang Bong Ahn, Dong Soo Han, Joong Ho Bae, Tae Jun Byun, Jong Pyo Kim, and Chang Soo Eun. The miss rate for colorectal adenoma determined by quality-adjusted, back-to-back colonoscopies. *Gut and liver*, 6(1):64, 2012.

[2] Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2):49–60, 1999.

[3] Irwan Bello, William Fedus, Xianzhi Du, Ekin Dogus Cubuk, Aravind Srinivas, Tsung-Yi Lin, Jonathon Shlens, and Barret Zoph. Revisiting resnets: Improved training and scaling strategies. *Advances in Neural Information Processing Systems*, 34:22614–22627, 2021.

[4] Carlo Biffi, Pietro Salvagnini, Nhan Ngo Dinh, Cesare Hassan, Prateek Sharma, and Andrea Cherubini. A novel ai device for real-time optical characterization of colorectal polyps. *NPJ digital medicine*, 5(1):84, 2022.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[6] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[7] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[8] Hala Fatima, Douglas K Rex, Richard Rothstein, Emad Rahmani, Omar Nehme, John Dewitt, Debra Helper, Arifa Toor, and Steven Bensen. Cecal insertion and withdrawal times with wide-angle versus standard colonoscopes: a randomized controlled trial. *Clinical Gastroenterology and Hepatology*, 6(1):109–114, 2008.

[9] Ori Kelner, Or Weinstein, Ehud Rivlin, and Roman Goldenberg. Motion-based weak supervision for video parsing with application to colonoscopy. *arXiv preprint arXiv:2210.10594*, 2022.

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] Jesse Lachter, Simon Christopher Schlachter, Robert Scooter Plowman, Roman Goldenberg, Yaffa Raz, Nadav Rabani, Natalie Aizenberg, Alain Suissa, and Ehud Rivlin. Novel artificial intelligence–enabled deep learning system to enhance adenoma detection: a prospective randomized controlled study. *iGIE*, 2023.

[12] George Leifman, Amit Aides, Tomer Golany, Daniel Freedman, and Ehud Rivlin. Pixel-accurate segmentation of surgical tools based on bounding box annotations. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 5096–5103. IEEE, 2022.

[13] Dan M Livovsky, Danny Veikherman, Tomer Golany, Amit Aides, Valentin Dashinsky, Nadav Rabani, David Ben Shimol, Yochai Blau, Liran Katzir, Ilan Shimshoni, et al. Detection of elusive polyps using a large-scale artificial intelligence system (with videos). *Gastrointestinal Endoscopy*, 94(6):1099–1109, 2021.

[14] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[15] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[16] Todd K Moon. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60, 1996.

[17] William Sanchez, Gavin C Harewood, and Bret T Petersen. Evaluation of polyp detection in relation to procedure time of screening or surveillance colonoscopy. *Official journal of the American College of Gastroenterology— ACG*, 99(10):1941–1945, 2004.

[18] Mandeep S Sawhney, Marcelo S Cury, Naama Neeman, Long H Ngo, Janet M Lewis, Ram Chuttani, Douglas K Pleskow, and Mark D Aronson. Effect of institution-wide policy of colonoscopy withdrawal time more than 7 minutes on polyp detection. *Gastroenterology*, 135(6):1892–1898, 2008.

[19] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

[20] Aasma Shaukat, Thomas S Rector, Timothy R Church, Frank A Lederle, Adam S Kim, Jeffery M Rank, and John I Allen. Longer withdrawal time is associated with a reduced incidence of interval cancer after screening colonoscopy. *Gastroenterology*, 149(4):952–957, 2015.

[21] Rebecca Shine, Andrew Bui, and Adele Burgess. Quality indicators in colonoscopy: an evolving paradigm. *ANZ journal of surgery*, 90(3):215–221, 2020.

[22] Dia T Simmons, Gavin C Harewood, Todd H Baron, Bret Thomas Petersen, Kenneth K Wang, F Boyd-Enders, and Beverly J Ott. Impact of endoscopist withdrawal speed on polyp yield: implications for optimal colonoscopy withdrawal time. *Alimentary pharmacology & therapeutics*, 24(6):965–971, 2006.

[23] Quirine EW van der Zander, Ramon M Schreuder, Roger Fonollà, Thom Scheeve, Fons van der Sommen, Bjorn Winkens, Patrick Aepli, Bu'Hussain Hayee, Andreas B Pischel, Milan Stefanovic, et al. Optical diagnosis of colorectal polyp images using a newly developed computer-aided diagnosis system (cadx) compared with intuitive optical diagnosis. *Endoscopy*, 53(12):1219–1226, 2021.

[24] Stephan R Vavricka, Michael C Sulz, Lukas Degen, Roman Rechner, Michael Manz, Luc Biedermann, Christoph Beglinger, Shajan Peter, Ekaterina Safroneeva, Gerhard Rogler, et al. Monitoring colonoscopy withdrawal time significantly improves the adenoma detection rate and the performance of endoscopists. *Endoscopy*, 48(03):256–262, 2016.

[25] Tao Yu, Ne Lin, Xu Zhang, Yanqi Pan, Huiyi Hu, Wenfang Zheng, Jiquan Liu, Weiling Hu, Huilong Duan, and Jianmin Si. An end-to-end tracking method for polyp detectors in colonoscopy videos. *Artificial Intelligence in Medicine*, 131:102363, 2022.

[26] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Birch: an efficient data clustering method for very large databases. *ACM sigmod record*, 25(2):103–114, 1996.