Semantic Parsing of Colonoscopy Videos with Multi-Label Temporal Networks

Ori Kelner Or Weinstein Ehud Rivlin Roman Goldenberg
Verily Life Sciences
{orikelner, orw, ehud, rgoldenberg}@google.com

Abstract

Following the successful debut of polyp detection and characterization, more advanced automation tools are being developed for colonoscopy. The new automation tasks, such as quality metrics or report generation, require understanding of the procedure flow that includes activities, events, anatomical landmarks, etc. In this work we present a method for automatic semantic parsing of colonoscopy videos. The method uses a novel DL multi-label temporal segmentation model trained in supervised and unsupervised regimes. We evaluate the accuracy of the method on a test set of over 300 annotated colonoscopy videos, and use ablation to explore the relative importance of various method's components.

1. Introduction

Optical colonoscopy is the standard of care procedure for colorectal cancer (CRC) screening. The primary target of screening colonoscopy is detecting polyps and preventively removing them. The first phase of the procedure (intubation) is inserting the endoscope all the way to the end the colon (cecum - see Fig. 1). This is followed by the second phase (withdrawal), when the endoscope is slowly pulled out, while examining the colon mucosa for the presence of lesions. For some symptomatic indications, it is recommended to go farther than the cecum, into the terminal ileum, which is the final part of the small intestine. When examining the rectum, it is recommended to deflect the endoscope camera backwards in a U-turn (rectal retroflextion maneuver) to allow better visualization of the distal rectum. When a polyp is detected, it is often resected or biopsied using tools inserted through the colonoscope instrument channel. In some cases, the colonoscope can be taken out of the body and re-inserted during the procedure.

The quality of the colonoscopy procedure is highly operator dependent and depends on the physician's skills, experience, fatigue, etc. To ensure high quality levels, professional societies recommend measuring and monitoring various quality metrics. For example, since not every procedure completes all recommended steps, cecal and ileum intubation rates and rectal retroflexion rate (percentage of procedures) are measured.

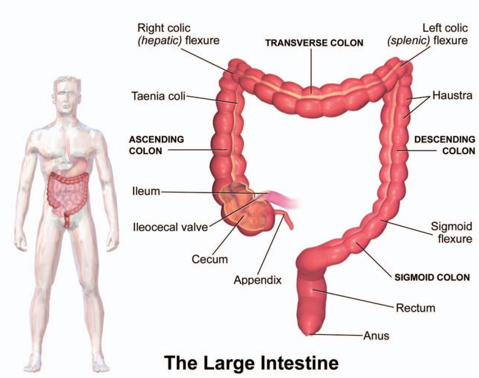

Figure 1. Colon Anatomy (from Wikipedia)

While the computer-aided tools for colonoscopy have been developed for years, only recently the first such tool - a polyp computed-aided detector (CADe) [15, 20, 11, 17, 18], became commercially available [20, 1]. This success triggered the development of other, more advanced computer-aided tools for colonoscopy, including automatic quality metrics, colonoscopy video annotation and retrieval, automatic report generation [20, 21, 16]. A common prerequisite for those tasks is the ability to parse a colonoscopy video into semantically meaningful parts, including activities/phases (e.g. intubation, withdrawal, polyp management, cleansing), events/key moments (e.g. polyp detected, retroflextion), anatomical landmarks (e.g. ileocecal valve) and segments (e.g. rectum, cecum, inside/outside the body, etc.).

The ability to automatically parse procedures has a lot of potential to improve current practice. Below we present some use-cases:

- Cecum detection enables straightforward calculation of withdrawal time [9], which is a standard quality metric that estimates the amount of time the physician is looking for polyps.
- Combined with polyp detection capabilities, the proposed method would enable localization of polyps

within the colon. This is crucial as some segments are more likely to contain precancerous polyps (adenomas).

- As mentioned earlier, semantic video parsing is a first step towards automatic report generation. This ability would potentially save time for physicians and could increase procedures volume.
- Detecting outside-body/inside-body allows removing the outside of the body video segments, which is a privacy requirement for using the colonoscopy videos for various research and clinical purposes.
- The ability to detect tools allows computing the so-called ”net withdrawal” time, which is the withdrawal time minus the time spent on polyp management. Net withdrawal time is a novel quality metric that might be better correlated to important clinical metrics such as adenoma detection rate (ADR).

Previous works on colonoscopy parsing were evaluated on significantly smaller datasets of around 20 videos [2, 3, 4]. [4], for example, uses a boundary detection algorithm that detects changes in blurriness and pixel intensity between the segments. A number of works [9, 10, 13] focus on parsing colonoscopy into withdrawal and intubation phases. Similar in spirit to our work are [7, 19] that perform temporal segmentation of laparoscopic videos into surgical phases. Specifically, [7] uses temporal networks on top of per frame features to detect surgical phases.

In what follows we present a method for colonoscopy video parsing and automatic detection of cecum, ileum, frames inside and outside of the body, rectal retroflexion, and use of surgical tools (see Fig. 6).

The main paper contributions are:

- A method to parse colonoscopy videos achieving 94.6% balanced accuracy on a large test set of 344 videos.
- An adaptation of a temporal convolution network to support multiple labels.
- A pseudo-labeling approach to increase the training set.

2. Methods

We follow a widely used two stage paradigm [6] for video parsing: first, extracting features from video frames using a single frame encoder, and then feeding them into a temporal classifier that captures high-level temporal patterns (see Fig. 2).

We propose several improvements to this straightforward approach, some of which are applicable to a wide range of scenarios and tasks, and some use the specific colonoscopy domain knowledge.

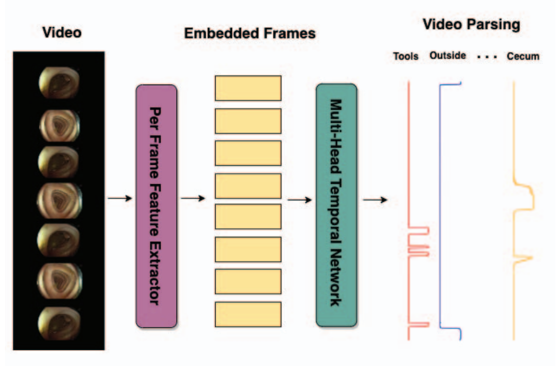

Figure 2. The two stage video parsing pipeline. The first stage is a single frame encoder. The second stage runs temporal convolution (MS-TCN, ASFormer) on frame embeddings to yield per-frame classifications.

2.1. Baseline Method

Our data consists of colonoscopy videos, annotated with the target labels (ileum, cecum, outside, inside, tools and rectal retroflextion) in the form of video segments, i.e. [start frame, end frame, label] triplets (see Section 3 for detailed dataset description). The straightforward approach for video parsing is to start with a pretrained (e.g. on ImageNet [5]) CNN, and use it as a single frame feature extractor for a temporal model. The temporal model, e.g. Multi-Stage Temporal Convolutional Network (MS-TCN) [6] is trained in a supervised way, using the annotated videos, to predict the labels (Fig. 2).

Note that our use case requires a multi-label approach. E.g., surgical tools may be used in any of the colon segments, hence the corresponding labels are non-mutually exclusive. Common architectures [6, 22] used for video segmentation do not support multiple labels natively, and we explain below how to adapt them.

2.2. Training Single-Frame Encoder with Key Frames

To improve over the baseline, we follow [7], where the frame encoder is pre-trained to predict the labels on a single frame. We use a shared CNN backbone, with multiple classification heads per each class (see Fig. 4). We sample random frames from labeled (annotated) segments, and train the model to predict the segment label. After the training is completed, we discard the classification heads, and use the shared backbone as the feature extractor for the temporal model, as before (Fig. 2).

Interestingly, with this approach, for most labels, we saw performance degradation (Table 2 - rows 1 vs. 2). We suspect the reason for this is that in colonoscopy many frames are not informative, e.g. due to camera motion blur, liquids, blocked view, etc. Such frames can be frequently found in annotated video segments. Hence, when training on frames randomly sampled from those segments, many of them are

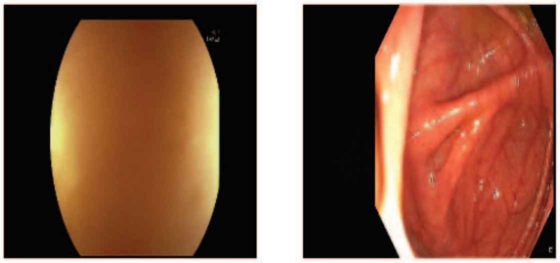

Figure 3. Left: A non-informative frame with a blocked field of view. Right: A key frame with a clear view of the triradiate fold.

not indicative of the segment label (see Fig. 3).

Instead, we suggest training the per-frame model with ”high-quality” frames that clearly represent the relevant label (Fig. 3). For example, for the cecum we leverage snapshots manually captured by gastroenterologist during the procedure to be included in the procedure report. Usually, in those snapshots the cecum landmarks (appendiceal orifice, triradiate fold and ileocecal valve) are clearly visible, hence providing a strong visual signal. We used the snapshots as positives frames to train the cecum head, and took random frames outside of the cecum segment as negatives. We see a significant improvement (Table 2 - row 3) of about 4% per-frame accuracy with this approach.

2.3. Pseudo Labels

Annotating large video datasets is both time and labor intensive. We can leverage a large pool of unlabeled data in order to increase the amount and diversity of the training samples. This might be important, as, for example, the ”outside” (of body) class could potentially contain a wide range of diverse scenes, so it’s essential to build a robust model for this task. In this section we focus on improving the single-frame encoder. The incorporation of temporal models follows in the next section.

Let $\mathcal{U} = \{(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), \ldots, (\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}$ be a small set of available labeled video **frames**, where $\mathbf{x}^{(i)}$ is the frame and $\mathbf{y}^{(i)}$ is its multi-label annotation. The labeled frames set $\mathcal{U}$ is composed of frames taken from the annotated video segments and snapshots of cecum landmarks and retroflexions, manually captured during the procedure (as explained in the Section 2.2).

Let K be the number of non-mutually exclusive labels, that is, $\mathbf{y}^{(i)} = (\mathbf{y}_k^{(i)}|_{k=1}^{K})$, where $\mathbf{y}_k^{(i)}$ is a one-hot vector for the k^{th} label of the i^{th} sample. In our case, the one-hot vectors correspond to the following non-exclusive labels: tools/no-tool, ileum/cecum/u-turn/other, and inside/outside.

Let $\mathcal{P} = \{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ be a large pool of available unlabeled **videos**. Let us denote the number of frames in video $\mathbf{v} \in \mathcal{P}$ by T.

2.3.1 Initial Supervised Model

We start by training a supervised model $\mathcal{F}$ using the annotated samples from $\mathcal{U}$. The model is composed of a shared feature extractor, followed by K different classification heads for each one of the K non-mutually exclusive labels (see Figure 4). We train the model to optimize the cross-entropy loss between the $\mathbf{y}_k^{(i)}$ and the model predictions $\widehat{\mathbf{y}}_k^{(i)}$:

$$L = \sum_i \sum_k CE(\mathbf{y}_k^{(i)}, \widehat{\mathbf{y}}_k^{(i)})$$

In principle, we don’t require each sample in $\mathcal{U}$ to have annotations for all K labels. If annotations for some labels are missing, we simply skip them in the summation over k.

2.3.2 Pseudo Labeling and Temporal Smoothing

To enrich the training set, we label the frames of unlabeled videos in $\mathcal{P}$ by applying the model $\mathcal{F}$ trained using the annotated data, as described in the previous section.

For every video $\mathbf{v} \in \mathcal{P}$, let $(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)})$ be the T frames of $\mathbf{v}$. Let $\widehat{\mathbf{y}}^{(t)}$ be the vector of predicted class probabilities for frame $\mathbf{x}^{(t)}$:

$$\widehat{\mathbf{y}}^{(t)} = \mathcal{F}(\mathbf{x}^{(t)}), t = 1..T$$

As before, $\widehat{\mathbf{y}}^{(t)} = (\widehat{\mathbf{y}}_k^{(t)}|_{k=1}^{K})$, where $\widehat{\mathbf{y}}_k^{(t)}$ is the vector of class probabilities for label k. In practice, as some of the tasks are difficult to predict from a single frame (cecum detection for example), we use the pseudo labeling approach only for inside-body/outside-body and tools detection tasks.

In order to reduce the pseudo-label noise, we smooth the class predictions along the temporal dimension t by a Gaussian kernel G_σ of size $2M + 1$ to yield

$$\widehat{\mathbf{y}}_k^{(t)} = (\widehat{\mathbf{y}}_k * G_\sigma)^{(t)} = \sum_{m=-M}^{M} \widehat{\mathbf{y}}_k^{(t-m)} G_\sigma[m]$$

2.3.3 Temporal Consistency Filtering

In order to further improve the quality of the pseudo-labeling, we leverage the prior domain knowledge about the temporal structure of colonoscopy procedures. We know that a procedure usually (but not always) follows the predefined sequence of phases:

1. Outside of the body
2. Inside the body
3. Outside of the body

We choose a very simple sanity check approach to discard videos with pseudo-labels that do not follow the very basic outside-inside-outside temporal pattern: Let us denote

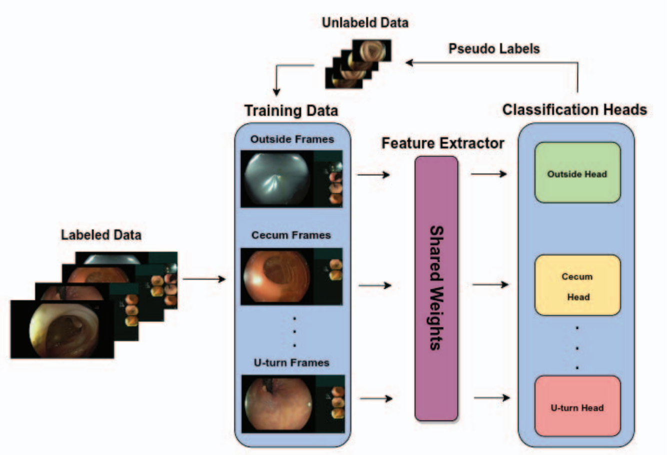

Figure 4. Pre-training of the feature extractor. We use a combination of labeled data, together with pseudo-labels as explained in Section 2.3. After the training is complete, we discard the classification heads and use the feature extractor to embed frames for the temporal network.

by $\widetilde{\mathbf{y}}^{(t)}_{\text{in/out}}$ the predicted inside/outside label class probability vector for frame t, where in/out is the index of the corresponding label. We require the start and end frames of the video to be outside of the body, and the middle frame of the video to be inside:

$$\arg\max_{l\in(0,1)} \widetilde{\mathbf{y}}^{(0)}_{\text{in/out}}[l] = \arg\max_{l\in(0,1)} \widetilde{\mathbf{y}}^{(T)}_{\text{in/out}}[l] = 1,$$
$$\arg\max_{l\in(0,1)} \widetilde{\mathbf{y}}^{(T/2)}_{\text{in/out}}[l] = 0.$$

In addition, tools almost never appear outside the body, as they are usually visible once the endoscope is inside the body and the physician is examining a polyp. Hence, videos in which we detect tools and outside the body over the same frame are also discarded. The reason we limit the temporal filtering to these simple heuristics is because complex temporal dependencies are introduced through a temporal network, as described in the following section. At this stage we are only interested to make sure the generated pseudo-labels are of reasonable quality, to reduce the label noise while training the single-frame encoder.

After discarding videos that don't meet these criteria, we re-train $\mathcal{F}$ to predict $\widetilde{\mathbf{y}}^{(i)}$ for pseudo-labeled videos $\mathbf{v} \in \mathcal{P}$ (for inside/outside and tools detection tasks), in addition to the annotated samples in $\mathcal{U}$. This way we significantly increase the size and the diversity of the training set.

2.4. Multi-Label Temporal Network

The main design improvement of MS-TCN [6] over the TCN [12] architecture, is the multi-stage approach. The first stage takes the frame embeddings and predicts a class for each frame (as in TCN), while the following stages "refine" those predictions. That is, the next stages are fed with the class-predictions of the previous stage, and the predictions of all stages equally contribute to the loss. A multi-stage design is also used by more modern action segmentation networks, such as the transformer-based ASFormer [22]. We cannot apply the MS-TCN approach in our case, as it does not support multi-label classification. To the best of our knowledge, there is no prior art that applies a multi-stage temporal network for a multi-label problem.

A naive adaptation of MS-TCN to deal with multi-label is to use separate networks for each label, at least from the 2nd stage on. This problem with this approach is that it does not allow any cross-talk between network signals corresponding to different labels. Our design allows all stages to exchange information related to different labels. This might be beneficial, as, for example, it is less common to see a tool in the ileum or outside of the body, and we would like the model to learn these priors and use them to refine predictions in later stages of the network. To enable this, we feed each stage with the concatenation of all class probabilities for all labels.

More formally, let $(\mathbf{x}^{(1)}, \ldots \mathbf{x}^{(T)})$ be the T frames of a video. Each frame $\mathbf{x}^{(t)}$ is labeled with K different labels $(\mathbf{y}^{(t)}_1, \ldots, \mathbf{y}^{(t)}_K)$. In our case $K = 3$, and $(\mathbf{y}^{(t)}_1, \mathbf{y}^{(t)}_2, \mathbf{y}^{(t)}_3)$ are the 2-,4- and 2-long one-hot vectors, corresponding to tool/no-tool, ileum /cecum/rectal-retroflextion/other, and inside/outside labels respectively. Let $\widehat{\mathbf{y}}^{(t)}_k$ be the vector of predicted class probabilities for k^{th} label of frame $\mathbf{x}^{(t)}$.

Our solution for the multi-label setup uses the per-label softmax applied to groups of logits corresponding to each label (see Fig. 5). The concatenated probabilities vector is then fed into the next stage.

Assume the network has S stages. Denote by $O^{(t)}_s$ the output of the s^{th} stage for the frame $\mathbf{x}^{(t)}$ (pre-softmax), and by I_k the indices of the logits relevant to the k^{th} label. Then the predicted vector of class probabilities for the k^{th} label, s^{th} stage, and t^{th} frame is

$$\widehat{\mathbf{y}}^{(t)}_{k,s} = \mathbf{softmax}(O^{(t)}_s[j] | j \in I_k). \tag{1}$$

The corresponding loss term, as defined in MS-TCN [6] is:

$$l^{(t)}_{k,s} = \frac{1}{T} CE(\widehat{\mathbf{y}}^{(t)}_{k,s}, \mathbf{y}^{(t)}_k) + \lambda \frac{1}{T|I_k|} ||\widehat{\mathbf{y}}^{(t-1)}_{k,s} - \widehat{\mathbf{y}}^{(t)}_{k,s}||^2_2, \tag{2}$$

where CE is the cross entropy loss between the ground truth for the t^{th} frame and the k^{th} label $\mathbf{y}^{(t)}_k$ and the prediction $\widehat{\mathbf{y}}^{(t)}_{k,s}$. The second loss term is a smoothing loss that encourages adjacent frames to have similar predictions. λ is a weighting factor. The final loss is computed over all stages, frames and labels:

$$loss = \sum_{s=0}^{S} \sum_{t=0}^{T} \sum_{k=0}^{K} w_k \cdot l^{(t)}_{k,s}, \tag{3}$$

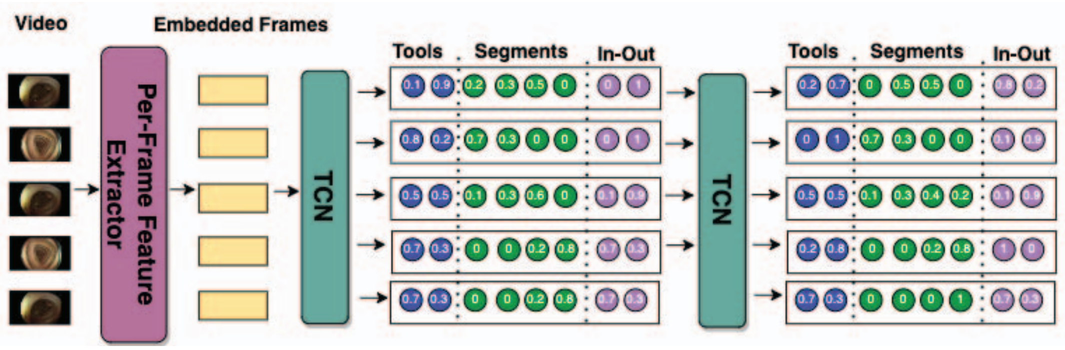

Figure 5. Multi-Label MS-TCN with two stages (the number of stages is a hyperparameter). Note that we apply the Softmax activation separately on the logits that correspond to the colon-segments, inside/outside and tools/no-tools.

Architectures	Avg. Accuracy	Ileum	Cecum	Rectal Retroflexion	Outside	Tool
ResNet, MS-TCN	90.4 ± 0.8	90.5 ± 0.7	89.6 ± 0.8	96.3 ± 1.0	$\mathbf{99.8} \pm 0.1$	$\mathbf{93.7} \pm 0.9$
ResNet, ASFormer	90.4 ± 0.9	88.8 ± 1.8	90.7 ± 0.4	97.1 ± 0.7	99.7 ± 0.2	91.0 ± 1.3
ConvNext, MS-TCN	94.1 ± 0.5	94.4 ± 0.9	$\mathbf{92.5} \pm 0.3$	98.7 ± 0.5	$\mathbf{99.8} \pm 0.1$	$\mathbf{93.7} \pm 0.4$
ConvNext, ASFormer	$\mathbf{94.6} \pm 0.5$	$\mathbf{96.1} \pm 0.5$	91.5 ± 0.7	$\mathbf{99.0} \pm 0.3$	$\mathbf{99.8} \pm 0.1$	92.1 ± 2.0

Table 1. Ablation study for network architectures. Average per-frame balanced classification accuracy over all labels, and for each label. Models trained with the ”key-frame” training scheme.

where w_k is a per-label weighting factor.

The proposed scheme enables multiple labels per frame, while introducing minimal changes to the original MS-TCN architecture.

3. Experiments and Results

3.1. Dataset

Our labeled data consists of 3,994 colonoscopy videos, recorded in 3 medical centers. We randomly split it into 344 videos for testing and 3,650 for training. We do not use a validation dataset as we do not perform hyper-paramter tuning in this paper, and leave this for future work. Instead, we use the commonly used parameters as described in Tables 4 and 5. We make use of labeled video segments annotated by experienced gastroenterologists (see Table 3). For each video, time segments were labeled, indicating when different colon-segments/tools appear, or whether the endoscope is outside of the body. For key-frames we use still-images of anatomical landmarks, which gastroenterologists manually captured during the colonoscopy procedure. In addition, we leverage the unlabeled set of 18,500 colonoscopy videos, by training the model on pseudo labels computed for these videos, as explained in Section 2.3.

All videos were standardized to 30 FPS, and had original resolution of 720P or 1080P (later we down-sample to 224 × 224 for training). The minimum bitrate used for compression was 12mbps, and the median procedure time is 11 minutes. Finally, the videos were captured by multiple endoscope types from 3 different manufactures: Olympus, Fuji and Pentax.

3.2. Accuracy Evaluation and Ablation Study

We preform an ablation study to better understand the role of different components. In particular, we compare the single frame encoder pre-trained on the Imagenet, with the one trained on random frames from annotated segments, and the one trained on key-frames. We measure the average per-frame classification accuracy on the test set, over all labels, and for each label. For the per-label results we use balanced accuracy (with equal weights for sensitivity and specificity) as the labels are heavily unbalanced. For each setup we ran 5 experiments and report the average and standard deviation. As can be seen in Table 2, our method with the key-frames training significantly outperforms the random sampling and the ImageNet baseline. As explained in Section 2.2, for most labels (ileum, cecum, rectal retroflextion and tools), the random sampling actually hurts the performance.

We also compare several network architectures: ResNet50 [8] and ConvNextBase [14] for feature extractors, and MS-TCN [6] and ASFromer [22] for temporal networks (Table 1). We notice a significant improvement with the larger and more modern ConvNextBase compared to ResNet50. On the other hand, the results of ASFromer and MS-TCN seem on-par. For the training settings and hyperparameters see Tables 4 and 5. Overall, as one can see, our proposed method achieves very high accuracy, reaching high 90s for most labels.

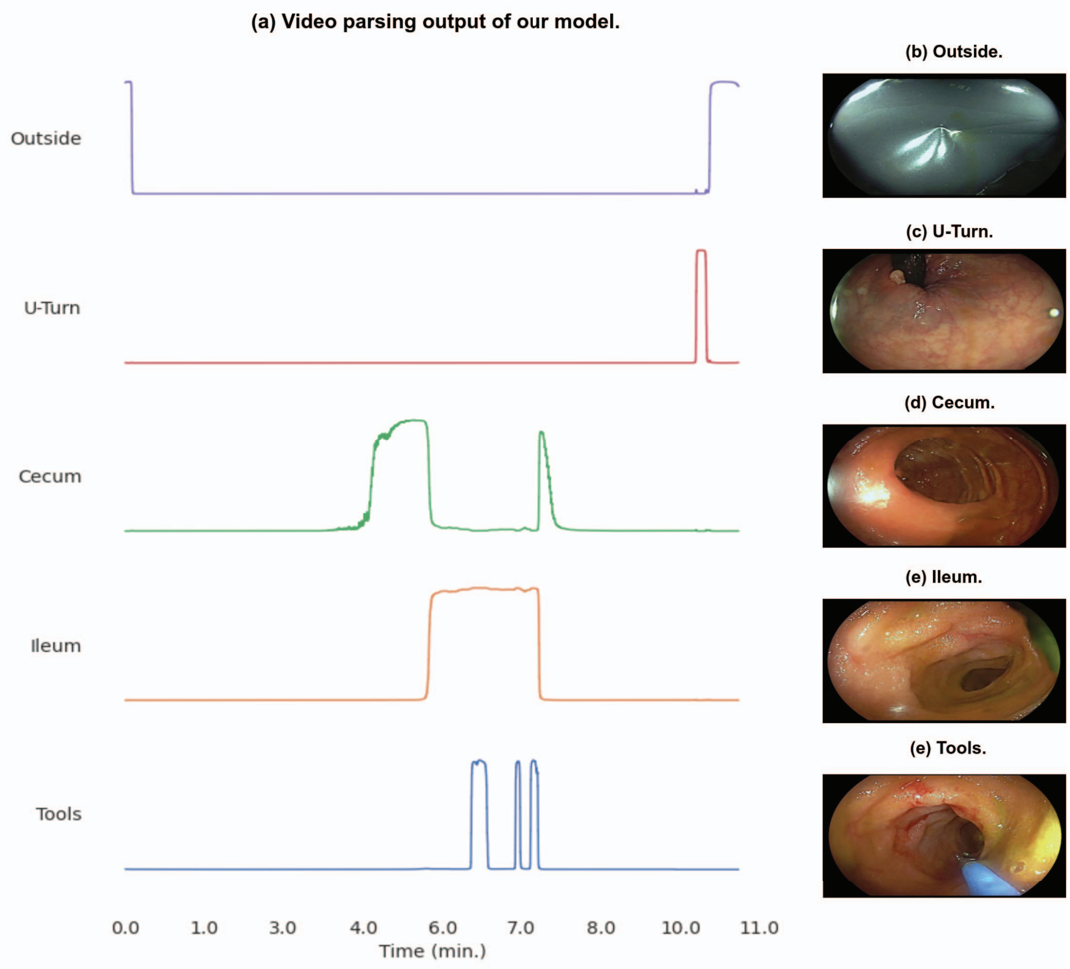

Figure 6. The output of our model: class probabilities over the course of the procedure and the corresponding video snapshots.

Method	**Avg. Accuracy**	**Ileum**	**Cecum**	**Rectal Retroflexion**	**Outside**	**Tool**
ImageNet pre-training	90.8 ± 1.0	92.3 ± 2.3	90.3 ± 0.8	94.9 ± 2.0	99.6 ± 0.3	90.2 ± 2.4
Classification on random labeled frames	87.2 ± 1.4	88.8 ± 2.7	89.5 ± 0.8	90.4 ± 3.5	99.7 ± 0.0	86.9 ± 2.4
Classification on key frames	$\mathbf{94.6} \pm 0.5$	$\mathbf{96.1} \pm 0.5$	$\mathbf{91.5} \pm 0.7$	$\mathbf{99.0} \pm 0.3$	$\mathbf{99.8} \pm 0.1$	$\mathbf{92.1} \pm 2.0$

Table 2. Ablation study for frame encoder training scheme. Average per-frame balanced classification accuracy over all labels, and for each label. Using ConvNextBase as the frame encoder and ASFormer as the temporal model.

Activity / Segment	**Snapshots**	**Annotated Segments**
Cecum	65K	2.5K
Rectal Retroflexion	20K	1.5K
Tools	-	14K
Terminal Ileum	-	1K
Inside and Outside	-	1.5K

Table 3. Annotations: number of annotated segments and stills for each label.

Parameter	**ResNet50v2**	**ConvNextBase**
Batch Size	64	64
Optimizer	Adam	Adam
Hardware	4 Tesla v100	4 Tesla v100
Num. of Param.	24M	88M
Gaussian kernel	$\sigma = 5, M = 10$	$\sigma = 5, M = 10$
Resolution	224×224	224×224

Table 4. Per-Frame Embedding Model training setup parameters.

4. Conclusions and Future Work

We presented a method for semantic parsing of colonoscopy videos. The proposed technique adapts the multi-stage temporal network (MS-TCN) to a multi-label scenario. To gain more accuracy, we improve the single frame feature extractor by training it on key-frames and

Parameter	MS-TCN	ASFormer
Batch Size	1	1
Optimizer	Adam	Adam
Stages	2	2
Layers per Stage	13	9
λ Smoothing loss factor	0.15	0.15
Hardware	1 Tesla v100	1 Tesla v100
Num. of Param.	0.5M	0.6M

Table 5. Temporal Model training setup parameters.

pseudo-labeling. The method is evaluated on hundreds of colonoscopies and demonstrates above 90% accuracy for all labels. Semantic parsing of colonoscopy videos enables a number of downstream applications, including quality metrics, video retrieval, and automatic report generation.

There are several promising directions for future work that can be based on this method and expand its capabilities. We plan adding additional colon segments such as the transverse, ascending, and descending colon. Another direction is automatic detection of various colonoscopy imaging modes including Narrow Band Imaging (NBI) and chromoendoscopy. Pursuing these avenues introduces more automation to colonoscopy, contributing to more accurate and efficient diagnosis and treatment.

References

[1] Markus Brand, Joel Troya, Adrian Krenzer, Costanza De Maria, Niklas Mehlhase, Sebastian Götze, Benjamin Walter, Alexander Meining, and Alexander Hann. Frame-by-frame analysis of a commercially available artificial intelligence polyp detection system in full-length colonoscopies. *Digestion*, 103(5):378–385, 2022.

[2] Yu Cao, Dalei Li, Wallapak Tavanapong, JungHwan Oh, Johnny Wong, and Piet C De Groen. Parsing and browsing tools for colonoscopy videos. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 844–851, 2004.

[3] Yu Cao, Wallapak Tavanapong, Kihwan Kim, Johnny Wong, JungHwan Oh, and Piet C De Groen. A framework for parsing colonoscopy videos for semantic units. In *2004 IEEE International Conference on Multimedia and Expo (ICME)(IEEE Cat. No. 04TH8763)*, volume 3, pages 1879–1882. IEEE, 2004.

[4] Yu Cao, Wallapak Tavanapong, Dalei Li, JungHwan Oh, Piet C De Groen, and Johnny Wong. A visual model approach for parsing colonoscopy videos. In *Image and Video Retrieval: Third International Conference, CIVR 2004, Dublin, Ireland, July 21-23, 2004. Proceedings 3*, pages 160–169. Springer, 2004.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] Yazan Abu Farha and Juergen Gall. MS-TCN: multi-stage temporal convolutional network for action segmentation. *CoRR*, abs/1903.01945, 2019.

[7] Tomer Golany, Amit Aides, Daniel Freedman, Nadav Rabani, Yun Liu, Ehud Rivlin, Greg Corrado, Yossi Matias, Wisam Khoury, Hanoch Kashtan, and Petachia Reissman. Artificial intelligence for phase recognition in complex laparoscopic cholecystectomy. *Surgical Endoscopy*, 36:1–9, 08 2022.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Liran Katzir, Danny Veikherman, Valentin Dashinsky, Roman Goldenberg, Ilan Shimshoni, Nadav Rabani, Regev Cohen, Ori Kelner, Ehud Rivlin, and Daniel Freedman. Estimating withdrawal time in colonoscopies. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pages 495–512. Springer, 2023.

[10] 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.

[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] Colin Lea, René Vidal, Austin Reiter, and Gregory D. Hager. Temporal convolutional networks: A unified approach to action segmentation. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops*, pages 47–54, Cham, 2016. Springer International Publishing.

[13] Ying Li, Alexander Ding, Yu Cao, Benyuan Liu, Shuijiao Chen, and Xiaowei Liu. Detection of endoscope withdrawal time in colonoscopy videos. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 67–74, 2021.

[14] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[15] 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.

[16] Bernd Münzer, Klaus Schoeffmann, and Laszlo Böszörmenyi. Content-based processing and analysis of endoscopic images and videos: A survey. *Multimedia Tools and Applications*, 77:1323–1362, 2018.

[17] Shimin Ou, Yixing Gao, Zebin Zhang, and Chenjian Shi. Polyp-yolov5-tiny: A lightweight model for real-time polyp detection. In *2021 IEEE 2nd International Conference on*

Information Technology, Big Data and Artificial Intelligence (ICIBA), volume 2, pages 1106–1111. IEEE, 2021.

[18] Ishak Pacal and Dervis Karaboga. A robust real-time deep learning based automatic polyp detection system. *Computers in Biology and Medicine*, 134:104519, 2021.

[19] Manfred Jürgen Primus, Klaus Schoeffmann, and Laszlo Böszörmenyi. Temporal segmentation of laparoscopic videos into surgical phases. In *2016 14th international workshop on content-based multimedia indexing (CBMI)*, pages 1–6. IEEE, 2016.

[20] Mahsa Taghiakbari, Yuichi Mori, and Daniel von Renteln. Artificial intelligence-assisted colonoscopy: A review of current state of practice and research. *World journal of gastroenterology*, 27(47):8103, 2021.

[21] Wallapak Tavanapong, JungHwan Oh, Michael A Riegler, Mohammed Khaleel, Bhuvan Mittal, and Piet C De Groen. Artificial intelligence for colonoscopy: Past, present, and future. *IEEE journal of biomedical and health informatics*, 26(8):3950–3965, 2022.

[22] Fangqiu Yi, Hongyu Wen, and Tingting Jiang. Asformer: Transformer for action segmentation. *arXiv preprint arXiv:2110.08568*, 2021.