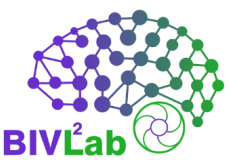
WEAKLY SUPERVISED DEEP REPRESENTATIONS TO SEGMENT COLORECTAL POLYPS IN CONTINUOUS COLONOSCOPY SEQUENCES

LINA MARCELA RUIZ GARCÍA

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2023



WEAKLY SUPERVISED DEEP REPRESENTATIONS TO SEGMENT COLORECTAL POLYPS IN CONTINUOUS COLONOSCOPY SEQUENCES

LINA MARCELA RUIZ GARCÍA

Research work in partial fulfillment of the requirements for the degree of:
Magíster en Ingeniería de Sistemas e Informática

Advisor:
Fabio Martínez Carrillo
Ph.D. in Systems and Computer Engineering

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2023

ACKNOWLEDGEMENTS

The author expresses her acknowledgment:

To God and to all the people who accompanied me in this project, who gave me their support and trained me both academically and personally. Those who helped to obtain this significant achievement in my professional life through their teachings and their example to follow.

I especially want to thank my advisor, Professor Fabio Martínez, whose training was invaluable. I admire and respect him for his dedication, understanding, and teaching. Thank you for all the learning you have given me since my undergraduate studies.

To my family and my boyfriend, Nicolás, for the advice and inspiration they have given me at every moment. For accompanying me in every goal I have set for myself, and of course, in the failures and successes. They are my motivation and my strength.

To my friends in the research group BIVL^{2}ab, for all the learning I got from each member, but especially to Alejandra and Edgar, for both the academic and personal support they gave me.

Finally, I would like to thank the UIS, which once again provided me with an excellent education. Thanks to "*Facultad de Físico-mecánicas*", "*Vicerrectoría académica*" and "*Movilidad UIS*", for supporting my participation in academic events.

Contents

List of Figures

List of Tables

pág.

LIST OF APPENDICES

pág.

ABSTRACT

Title: Weakly supervised deep representations to segment colorectal polyps in continuous colonoscopy sequences [*]

Author: Lina Marcela Ruiz García [**]

Keywords: Colorectal cancer, weakly supervised models, saliency maps, polyp segmentation, colonoscopy continuous sequences.

Description: Colorectal cancer is the third most commonly diagnosed cancer worldwide. Polyps are considered the main biomarkers of this cancer, being typically observed from optical colonoscopies. Nonetheless, the detection and shape characterization is challenging, even for expert gastroenterologists, because of shape and appearance variability, intestinal tract artifacts, and noise observations from colonoscopies. In fact, clinical studies revealed polyp loss of up to 26% during a clinical routine, impacting the early diagnosis and patient treatment. Some computational approaches have evidenced remarked support for polyp characterization but rely dependent on supervised representations, working on scenarios with relatively well-defined polyp presence. Far from such an assumption, colonoscopies in real scenarios are long sequences where polyps are isolated and scarce observations concerning the intestinal tract. This work introduces a multi-head cross-attention strategy to segment polyps under a weakly supervised scheme, including unlabeled background frames. While the cross-attention mechanism recovers polyp patterns by learning the non-local relationship of pixels through dilated convolutions, the minimization rule learns to differentiate between frames with polyp and background. The proposed approach is validated in a retrospective study that includes 40 long colonoscopy sequences (on average 15.000 frames per video). Despite the approximation of state-of-the-art, this dataset represents the first effort to approximate the segmentation problem in real scenarios. The proposed method achieves a precision of 70%, and a recall of 75% in long sequences, showing remarkable performance. Also, the proposed approach was validated on five public datasets outperforming the state-of-the-art with a precision of 92% (ASU-Mayo) and 96% (CVC-Video).

---

[*] Degree work

[**] Facultad de Ingenierías Fisicomecánicas. Escuela de Ingeniería de Sistemas e Informática. Advisor: Fabio Martínez Carrillo, Ph.D.

RESUMEN

Título: Representaciones profundas débilmente supervisadas para la segmentación de pólipos colorrectales en secuencias continuas de colonoscopia [*]

Autor: Lina Marcela Ruiz García [**]

Palabras clave: Cáncer colorrectal, modelos débilmente supervisados, mapas de saliencia, segmentación de pólipos, secuencias continuas de colonoscopia.

Descripción: El cáncer colorrectal es el tercer cáncer más diagnosticado a nivel mundial. Los pólipos se consideran los principales biomarcadores de este cáncer, observándose a partir de colonoscopias. Sin embargo, la detección y caracterización de estas lesiones es un reto, incluso para los gastroenterólogos expertos, debido a la variabilidad en forma y apariencia, los artefactos del tracto intestinal, y las observaciones ruidosas de las colonoscopias. De hecho, estudios clínicos revelan una pérdida de pólipos de hasta el 26% durante una rutina clínica, lo que repercute en el diagnóstico precoz y el tratamiento de los pacientes. Algunos enfoques computacionales han apoyado la caracterización de pólipos, pero dependen de representaciones supervisadas, trabajando en secuencias con pólipos relativamente bien definidos. Alejadas de tal suposición, las colonoscopias en escenarios reales son secuencias largas donde los pólipos son observaciones aisladas y escasas en cuanto al tracto intestinal. Este trabajo propone una estrategia de atención cruzada multi-cabeza para segmentar pólipos bajo un esquema débilmente supervisado, incluyendo fotogramas de fondo. Mientras que el mecanismo de atención extrae patrones del pólipo aprendiendo relaciones no locales de los píxeles a través de convoluciones dilatadas, la regla de minimización diferencia entre imágenes con pólipo y fondo. La validación del método propuesto se realizó en un estudio retrospectivo que incluye 40 colonoscopias ($\sim$ 15.000 fotogramas por video). A pesar de la aproximación del estado del arte, este conjunto de datos representa el primer esfuerzo para aproximar la segmentación en escenarios reales. El método propuesto obtiene 70% de precisión y 75% de sensibilidad en secuencias largas, mostrando un rendimiento destacable. Asímismo, el método fue validado en conjuntos de datos públicos superando al estado del arte con una precisión de 92% (ASU-Mayo) y 96% (CVC-Video).

---

[*] Trabajo de grado

[**] Facultad de Ingenierías Fisicomecánicas. Escuela de Ingeniería de Sistemas e Informática. Director: Fabio Martínez Carrillo, Ph.D.

INTRODUCTION

Colorectal cancer (CRC) is the third most common cancer in the world, reporting an incidence of around $1,9$ million diagnostic cases and a dramatic mortality rate of 900.000 deaths[1]. Polyps, as the main CRC biomarkers, are abnormal masses that grow in the lining of the colon or rectum. Colonoscopy is nowadays the standard test to observe, search for and characterize this biomarker. During this procedure, it is possible to biopsy the masses and detect and remove suspicious areas.

A critical point during colonoscopies is the early detection of polyps because their characterization is only performed by visual inspection based on geometric and textural patterns[2]. In particular, polyps smaller than 5 mm are associated with early stages or may be considered benign lesions. If these lesions are missed during the procedure, their size will progressively increase to the point of being considered malignant[3]. Hence, the early detection of such masses enables an effective clinical extraction procedure, further analysis, and early treatments, increasing the survival rate of up to 93%. Contrarily, the larger polyps (around 20 mm or more) are classified in the advanced stages of cancer, and the survival rate has a dramatic index rate of 8%[4].

Effective polyp detection requires detailed observations by experts, taking an average

---

[1] Jacques FERLAY et al. “Cancer statistics for the year 2020: An overview”. In: *International Journal of Cancer* 149.4 (2021), pp. 778–789.

[2] Sae HWANG et al. “Polyp Detection in Colonoscopy Video using Elliptical Shape Feature”. In: *2007 IEEE International Conference on Image Processing*. Vol. 2. 2007, pp. II–465.

[3] Norton GREENBERGER et al. *Current Diagnosis and Treatment: Gastroenterology, Hepatology, and Endoscopy*. 3rd ed. Mc Graw Hill, 2016.

[4] Eduardo PÉREZ. *Gastroenterología*. 1st ed. McGraw Hill Mexico, 2012.

of 30 minutes per exam. Besides, polyps may present high appearance variability due to constant changes in illumination, regions in the intestinal tract, and camera position, difficulting the characterization. Several studies have reported that during the clinical procedure, 6 to 25% of polyps may be missed due to factors associated with the expertise of the gastroenterologist, the location of the lesion, the preparation of the patient, and the textural similarity that this mass often has with the intestinal tract[5].

Hence, computational methods have been adapted to support the detection, segmentation, and characterization of polyps during colonoscopy procedures. Currently, these strategies are mainly based on deep representations, for instance, using as backbone U-nets, integrating attention modules to highlight polyp features, and implementing deconvolution modules to recover a proper polyp shape[6] [7] [8]. These strategies have demonstrated remarkable capabilities to delineate polyps in complex scenarios of the intestinal tract that share appearance features with abnormal masses. Besides, these methods are relatively well adapted to the variability of shape, color, and texture patterns for coding polyps. Nevertheless, such deep representations are based on cropped datasets focused only on polyps observations. It should be noted that polyps only represent less than 10% of the frames of atypical colonoscopies. So, the validation is far from realistic scenarios.

---

[5] Quentin ANGERMANN, Aymeric HISTACE, and Olivier ROMAIN. "Active learning for real time detection of polyps in videocolonoscopy". In: *Procedia Computer Science* 90 (2016), pp. 182–187.

[6] Olaf RONNEBERGER, Philipp FISCHER, and Thomas BROX. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer. 2015, pp. 234–241.

[7] Deng-Ping FAN et al. "Pranet: Parallel reverse attention network for polyp segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 263–273.

[8] Chien-Hsiang HUANG, Hung-Yu WU, and Youn-Long LIN. "Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps". In: *arXiv preprint arXiv:2101.07172* (2021).

Such approaches may fail in polyp detection over long sequences, generating a significantly high amount of false positive polyps, which may increase delays in the colonoscopy analysis. Additionally, in colonoscopies, the texture and appearance patterns can change dramatically, which may be unacceptable for deep representations.

This work introduces a novel polyp segmentation architecture, adjusted from a weakly supervised strategy, approaching non-labeled tract information to address the detection and characterization of abnormal masses during long colonoscopy sequences. The main contributions are described in the following items:

- A robust deep model that integrates cross-attention modules to recognize polyps from colonoscopy sequences. This attention module operates under dilated convolutional operation, which enables a coarse and non-linear representation of regions of interest.
- A weakly supervised training scheme that integrates polyp annotations and intestinal tract information without polyps. Thus, a special logistic minimization rule is introduced to weight-correct polyp segmentation and penalizes false positive detection.
- A novel dataset including long colonoscopy sequences, with an average length of 15.000 frames. A total of 40 long video sequences were selected to deal with the polyp detection and segmentation problem, resulting in a closer approximation to real scenarios.

1. FUNDAMENTALS

1.1. Colorectal cancer

Colorectal cancer (CRC) is the third most aggressive and the second most deadly cancer worldwide[9]. Polyps, as the main CRC biomarkers, are masses of tissue that grow into the intestinal tract, typically observed from standard colonoscopies. These polyps are classified according to their morphology in non-adenomas, commonly less than 5 mm, with a similar texture to the intestinal tract. Likewise, neoplastic adenomas are prominent lesions with pronounced textural features associated with a malignant prognosis[10]. In fact, around 60% of polyps constitute potential adenoma samples, and only 39% of these neoplasms are diagnosed at early stages[11]. These facts are particularly dramatic since early detection represents a survival rate of up to 90%[12].

Polyp morphology and shape are classified using different protocols, such as the Paris classification, which divides polyps into neoplasms. These neoplasms are further categorized as polypoid or non-polypoid. A polypoid has a pedunculate tree-like stem (0-Ip) and sessile protrusion (0-Is) above the submucosa. The non-polypoid neoplastic lesions include slightly elevated (0-IIa), flat (0-IIb), slightly depressed (0-IIc), and excavated (0-III)

---

9 FERLAY et al., “Cancer statistics for the year 2020: An overview”.

10 Shinji TANAKA et al. “Evidence-based clinical practice guidelines for management of colorectal polyps”. In: *Journal of Gastroenterology* 56 (2021), pp. 323–335.

11 Haoqi GAO and Koichi OGAWARA. “Adaptive data generation and bidirectional mapping for polyp images”. In: *IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*. IEEE. 2020, pp. 1–6.

12 Juan ORTEGA-MORÁN et al. “Medical needs related to the endoscopic technology and colonoscopy for colorectal cancer diagnosis”. In: *BMC cancer* 21.1 (2021), pp. 1–12.

lesions attached to the submucosa[13]. All these polyp types and sub-types are illustrated in Figure 1.

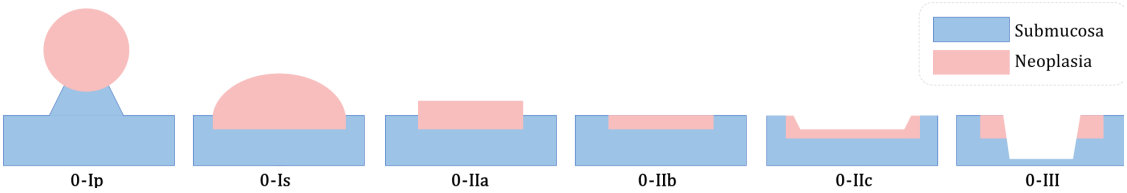


Figure 1. Paris classification type 0. In blue is the submucosa, and in red is the shape of the polyp for each class.

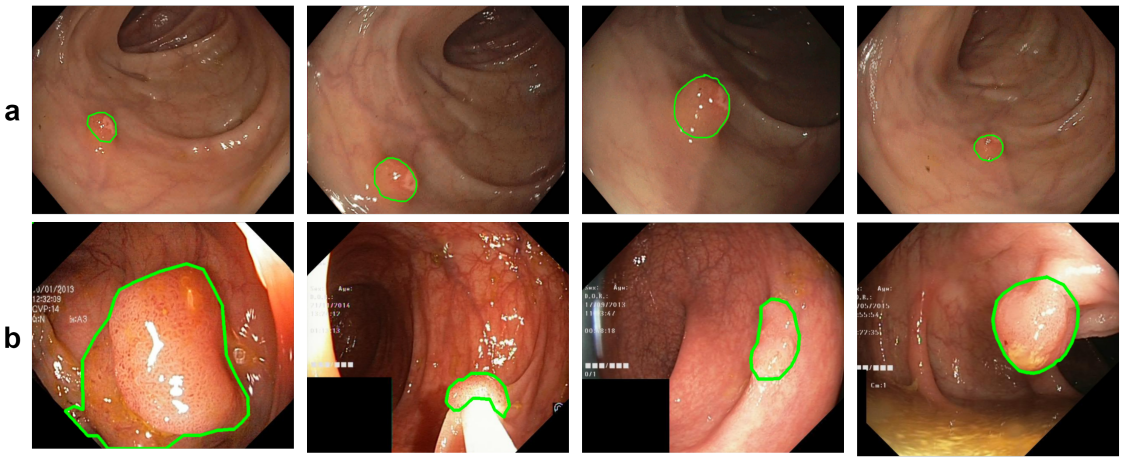
Colonoscopy is an invasive procedure considered the gold standard for detecting and removing colorectal lesions or abnormalities associated with CRC by visual inspection of the intestinal tract. Patients should have a previous preparation that allows exploration of the intestinal tract in clean and adequate conditions. However, polyp detection and characterization are challenging because their morphology varies among patients, abrupt camera movements, and a successful examination depends, among others, on proper patient preparation and the expertise of a gastroenterologist. Moreover, the textural similarity of polyps regarding the intestinal tract makes segmentation considerably complex, resulting in missed polyp detection between 6-25%. These dramatic statistics directly impact the early diagnosis and treatment of the disease[14]. Figure 2 shows polyp samples observed in colonoscopy sequences.

---

[13] R LAMBERT. "The Paris endoscopic classification of superficial neoplastic lesions: esophagus, stomach, and colon: November 30 to December 1, 2002". In: *Gastrointestinal Endoscopy* 58 (2003), S3–S43.

[14] ANGERMANN, HISTACE, and ROMAIN, "Active learning for real time detection of polyps in video-colonoscopy".

Figure 2. Typical polyps observations from colonoscopies. The green region delineates the polyp region. (a) The same polyp but at various perspectives taken during a colonoscopy sequence. (b) Different polyp examples at several procedures.



1.2. Image segmentation using deep learning

Several computational methods have been proposed to address image segmentation in diverse clinical domains. Recently, segmentation is mainly approximated from a machine learning scheme by classifying each pixel with an associated class. These strategies have been implemented from deep convolutional frameworks, following training schemes as supervised and weakly supervised learning, described as follows:

1.2.1. Convolutional networks have been adapted to perform per-pixel classification and approximate the segmentation task through sub-sampling and up-sampling operations[15]. In these models, the representation of the image is embedded into a compact vector to extract the principal features of the region of interest. This first scheme of encoding is named *encoder*. The resultant vector from the encoding is a hidden latent vector that learns the principal relationships regarding a segmentation task. Then, this vector is de-

[15] Jonathan LONG, Evan SHELHAMER, and Trevor DARRELL. "Fully convolutional networks for semantic segmentation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 3431–3440.

coded by disaggregation operations, obtaining the segmentation of the image according to the considered classes.

The U-net is the most representative convolutional architecture implemented for object segmentation[16]. In such a network, the encoder collects the contextual information and compresses the input image using pooling layers. The decoder expands the embedding representation to assign a class to each pixel, achieved by the up-sampling operations. A main contribution of U-net is the transferring context information mechanism from the encoder to the decoder to guided segmentation[17], as observed in Figure 3.

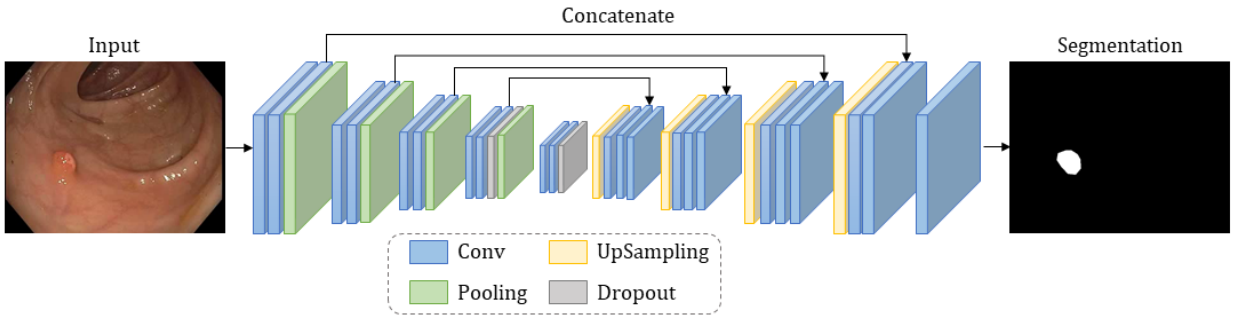


Figure 3. Convolutional approach based on the encoder-decoder strategy.

1.2.2. Dilated convolutions Convolutional architectures have also integrated different schemes in the encoding and decoding stages to improve the characterization of the region of interest. Such mechanisms have achieved more robust deconvolutions, thus generating a more accurate segmentation. The dilated convolutions have been widely explored for the advantages of extracting long dependencies without increasing the number of parameters. These convolutions allowed modeling a multi-scale version of the receptive fields, capturing non-local relationships through multi-branch convolutional block denom-

---

[16] RONNEBERGER, FISCHER, and BROX, "U-net: Convolutional networks for biomedical image segmentation".

[17] Liangliang LIU et al. "A survey on U-shaped networks in medical image segmentations". In: *Neurocomputing* 409 (2020), pp. 244–258.

inated Receptive Field Block (RFB)[18]. Each branch of the RFB is a convolutional layer of defined kernel size, followed by a pooling or convolution layer with the respective dilation parameter. The feature maps obtained from all branches are then concatenated in a block. In most cases, these feature maps are computed from the last convolutional layers of the architecture to reinforce the features learned by the network. Figure 4 represents the computation of RFB from a feature map. For instance, the dilated convolutions have been integrated into a Resnet-50 to compute contextual information from the image[19].

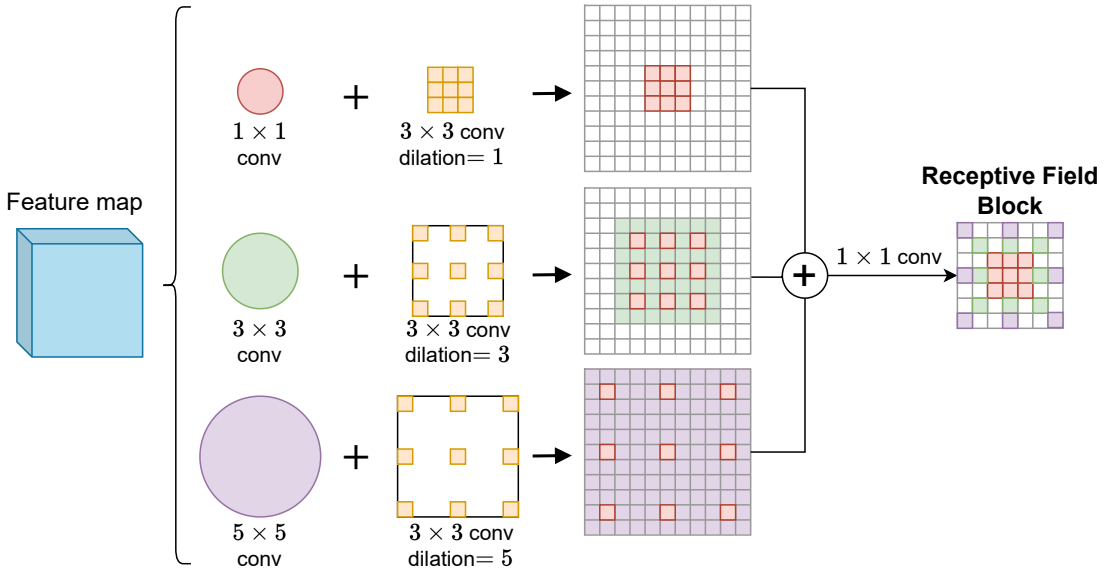


Figure 4. Calculation of RFB from a feature map. Several convolutions with different kernels and dilations are concatenated in a block using a 1×1 convolution.

1.2.3. Attention-based architectures have recently gained attention as mechanisms to recover non-local relationships between specific input regions. Instead of learning patterns from all pixels, the attention mechanism discovers the most relevant information

---

[18] Songtao LIU and Di HUANG. "Receptive field block net for accurate and fast object detection". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 385–400.

[19] Yunbo GUO, Jorge BERNAL, and Bogdan J. MATUSZEWSKI. "Polyp Segmentation with Fully Convolutional Deep Neural Networks Extended Evaluation Study". In: *Journal of Imaging* 6.7 (2020).

related to the class of interest and thus focuses on that region.

The origin of these attention mechanisms is natural language processing for machine translation using sequence-to-sequence recurrent networks[20]. Today, these mechanisms have been extended to several applications, including image representation, video analysis, and, of course, object segmentation. For image representation, these attention mechanisms have been modeled from convolutional architectures, starting the processing from a bank of deep features (f) and allowing calculating spatial and non-local information[21]. The simplest mechanism is self-attention which defines three primary branches to extract non-linear projections. Specifically, key $(k = W_k f)$, query $(q = W_q f)$, and value $(v = W_v f)$, where $W_k, W_q,$ and W_v represent the weight matrices of each projection. These projections can be 2D convolutions or 1×1 convolutions. A similarity operation is performed between the key and query branches to obtain an attention matrix that encodes the non-local correlation across pixels. This matrix improves the projections acquired by the value branch through matrix multiplication. This process is defined as

$$\text{Attention}(k, q, v) = \text{softmax}\left(\frac{qk^T}{\sqrt{d_k}}\right) v \tag{1}$$

where d_k, d_q, and d_v represents the dimensions of k, q and v, respectively. More robust approaches, instead of extracting a single attention module, compute attention mechanisms at the same level *i.e.* in parallel, which improves the characterization of the object of interest, learning distinct properties by applying different linear projections. The compu-

---

[20] Dzmitry BAHDANAU, Kyunghyun CHO, and Yoshua BENGIO. "Neural machine translation by jointly learning to align and translate". In: 2014.

[21] Long CHEN et al. "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 5659–5667.

tation of these multiple attention mechanisms is named the multi-head attention defined as $\text{Multi-head} = \text{Concat}(head_1, \ldots, head_n)$, being the concatenation of multiple attentions heads $(head_n)$, where n is the number of extracted attention modules. Figure 5 represents this procedure defined from a set of keys (K), queries (Q), and values (V)[22].

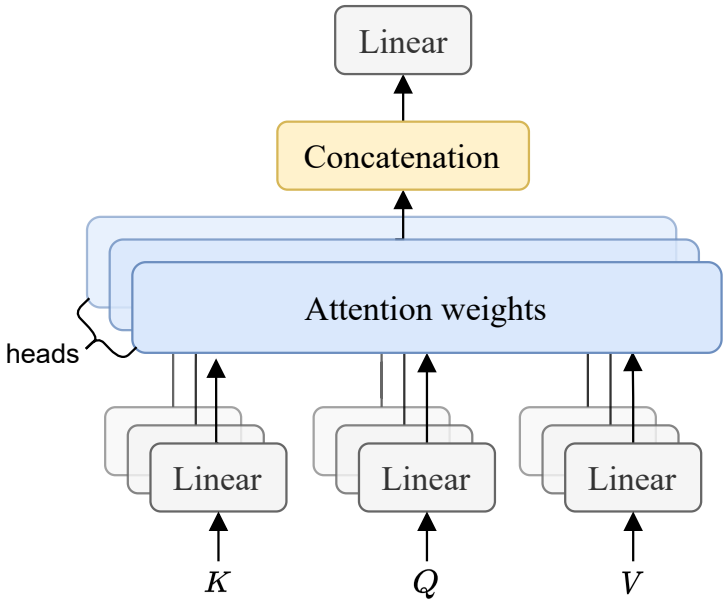


Figure 5. Multi-head attention mechanism with several heads extracted in parallel.

Weakly supervised learning aims to perform per-pixel prediction without relying on high-quality annotations. This type of learning is motivated by the training with few labels due to the effort required and the bias associated with annotations. In the literature, the related methodologies typically use labels as (1) bounding boxes, delimiting the region of interest by coordinates; (2) scribbles, which draw a line over each object[23]; (3) points, that define the center of the object; and (4) the image-level classification.

---

[22] Ashish VASWANI et al. "Attention is all you need". In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.

[23] Di LIN et al. "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 3159–3167.

2. PREVIOUS WORK: POLYP SEGMENTATION AND LOCALIZATION

Polyp localization and representation have been exhaustively explored to mitigate subjectivity in the diagnosis and miss-detection. The proposed approaches have explored strategies based on hand-crafted features, encoder-decoder methods, and attention-based architectures. Figure 6 illustrates the principal schemes adopted for polyp segmentation and localization.

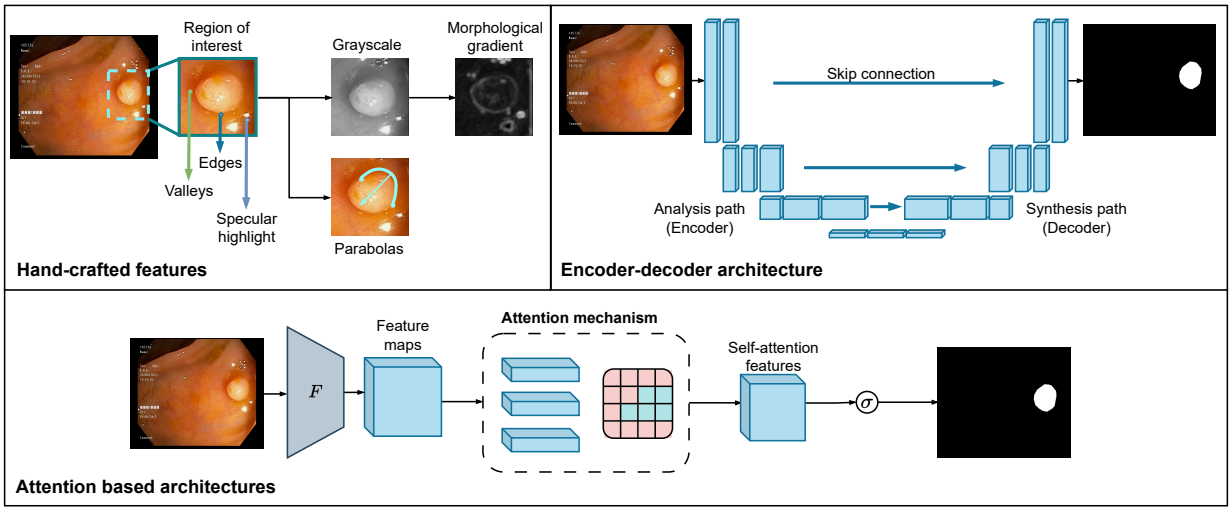


Figure 6. State-of-the-art methods to characterize polyps in colonoscopy procedures.

The first family of approaches has proposed image analysis descriptors capturing edges and approximating the masses to elliptical shape primitives[24]. Additionally, hand-crafted strategies included energy gradient maps, from which polyps may emerge as the valleys generated by the incidence of light[25]. Nonetheless, these strategies are based on specific

---

[24] HWANG et al., "Polyp Detection in Colonoscopy Video using Elliptical Shape Feature".

[25] Jorge BERNAL et al. "Polyp segmentation method in colonoscopy videos by means of MSA-DOVA energy maps calculation". In: *Clinical Image-Based Procedures. Translational Research in Medical Imaging*. Springer. 2014, pp. 41–49.

engineering of features that result in sensible scenario variations, making it challenging to generalize polyp morphology during clinical procedures.

Deep representations have demonstrated remarkable capabilities to characterize polyps from the shape and textural patterns, dealing with scene variations captured during intestinal track analysis. In such a line of work, a second family of approaches was implemented from encoder-decoder architectures, representing a robust alternative to obtain segmentation maps from deconvolution paths or implement a skip-layer structure to improve representation from high- and low-level features[26] [27]. Along the same line of thought, some approaches have implemented U-net architectures, adapting dilated convolutions to retrieve polyp shapes in relatively controlled scenarios over temporal sequences[28]. Similarly, an extended version of Res-Unet++ implementing Conditional Random Field (CRF) and Test-Time Augmentation (TTA) is proposed to improve the polyp representation[29]. This architecture uses squeeze and excitation blocks, which allow the extraction of global features, while the CRF statistical modeling captures context information. Despite advances in these representations, receptive fields adjusted with only polyp features may result in challenging and high variables concerning polyp nature. Also, these modules only explore such mechanism into convolutional layers, which increase locality but remains limited to computing global descriptors that consider distant polyp features. Other

---

[26] Patrick. BRANDAO et al. “Fully convolutional neural networks for polyp segmentation in colonoscopy”. In: *Medical Imaging 2017: Computer-Aided Diagnosis*. Vol. 10134. SPIE. 2017, pp. 101–107.

[27] Qiaoliang LI et al. “Colorectal polyp segmentation using a fully convolutional neural network”. In: *2017 10th International Congress on Image and Signal Processing, Biomedical Engineering and Informatics (CISP-BMEI)*. IEEE. 2017, pp. 1–5.

[28] Xinzi SUN et al. “Colorectal polyp segmentation by u-net with dilation convolution”. In: *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. IEEE. 2019, pp. 851–858.

[29] Debesh JHA et al. “A Comprehensive Study on Colorectal Polyp Segmentation With ResUNet++, Conditional Random Field and Test-Time Augmentation”. In: *IEEE Journal of Biomedical and Health Informatics* 25.6 (2021), pp. 2029–2040.

approaches have explored methods to improve polyp segmentation but from analyzing binary masks. For example, data augmentation based on meta-learning mixtures and implementation of confidence-aware resampling methods to address both missing annotated polyp segmentation datasets and the problem of high polyp variability shows stable performance on colonoscopy and WCE datasets, or implementation of 2D Gaussian masks instead of binary masks to minimize the encoder-decoder architecture evidences remarkable results on small and flat polyps[30] [31].

More recently, in a third family of strategies, the attention modules have recently been included in polyp segmentation nets to provide non-linear feature relationships. These modules have shown robust appearance representations to discriminate between these masses and intestinal folds. The PraNet architecture integrates three parallel reverse attention modules at the end of the convolutional backbone[32]. These reverse attention modules force deep features to delete the background by multiplying an estimated polyp mask. This approach, however, may be sensible to segmentation masks (attention weights), filtering out relevant regions into deep feature layers. Complementary, the HarDNet-MSeg architecture proposed a convolutional representation with receptive field blocks and without attention, achieving remarkable polyp segmentation results[33]. These receptive blocks include dilated convolutions to supply non-local spatial dependencies, but these modules may have strong dependencies of parameters. Other approaches have proposed a two-stage method to detect and segment polyps. A Faster R-CNN for the polyp proposal stage

---

[30] Xiaoqing GUO et al. “Non-equivalent images and pixels: Confidence-aware resampling with meta-learning mixup for polyp segmentation”. In: *Medical Image Analysis* 78 (2022), p. 102394.

[31] Hemin Ali QADIR et al. “Toward real-time polyp detection using fully CNNs for 2D Gaussian shapes prediction”. In: *Medical Image Analysis* 68 (2021), p. 101897.

[32] FAN et al., “Pranet: Parallel reverse attention network for polyp segmentation”.

[33] HUANG, WU, and LIN, “Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps”.

and a Fully Convolutional Network to the segmentation[34]. Similarly, the ColonSegNet, an encoder-decoder architecture for polyp detection, localization, and segmentation using residual blocks with squeeze and excitation modules[35]. Nevertheless, these results are achieved on relatively controlled sequences and may fail in typical scenarios with abrupt camera movements of hidden abnormal masses.

Polyp-PVT (Pyramid Vision Transformers) like architectures have been adapted to polyp characterization tasks[36]. This architecture can recover high- and low-level features aggregated in a similarity module to compute the most relevant relationships among characteristics following a global attention modeling. Also, the SwinE-Net implemented the EfficientNet and Vision Transform (ViT)-based Swin Transformer to compute spatial and semantic features at a patch- and global-level[37]. This architecture uses a multi-dilation convolutional block to refine multilevel feature maps and multiple feature aggregation modules to compute similarities. These architectures have demonstrated a better representation of polyps in public datasets. However, learning highly variable intestinal backgrounds may be essential to generalizing polyp features, avoiding false positives usually generated by water bubbles, specular reflections, or intestinal folds. Some approaches have implemented semi-supervised strategies to train a U-net backbone that refines polyp segmentation. Such an approach represents an effort to learn from scarce labels, but their modeling re-

---

[34] Xiao JIA et al. "Automatic Polyp Recognition in Colonoscopy Images Using Deep Learning and Two-Stage Pyramidal Feature Prediction". In: *IEEE Transactions on Automation Science and Engineering* 17.3 (2020), pp. 1570–1584.

[35] Debesh JHA et al. "Real-time polyp detection, localization and segmentation in colonoscopy using deep learning". In: *IEEE Access* 9 (2021), pp. 40496–40510.

[36] Bo DONG et al. "Polyp-PVT: Polyp Segmentation with Pyramid Vision Transformers". In: *arXiv e-prints, arXiv-2108* (2021).

[37] Kyeong-Beom PARK and Jae Yeol LEE. "SwinE-Net: Hybrid deep learning approach to novel polyp segmentation using convolutional neural network and Swin Transformer". In: *Journal of Computational Design and Engineering* 9.2 (Apr. 2022), pp. 616–632. ISSN: 2288-5048.

mains limited to controlled scenarios with only labeled polyp frames[38]. Other approaches implemented a hybrid polyp segmentation using a 2D/3D encoder-decoder network to address the problem of static colonoscopy images and take advantage of the spatial and temporal correlations obtained through 2D and 3D convolutional layers, respectively[39].

[38] Siwei CHEN, Gregor URBAN, and Pierre BALDI. "Weakly Supervised Polyp Segmentation in Colonoscopy Images Using Deep Neural Networks". In: *Journal of Imaging* 8.5 (2022). ISSN: 2313-433X.

[39] Juana González-Bueno PUYAL et al. "Polyp detection on video colonoscopy using a hybrid 2d/3d cnn". In: *Medical Image Analysis* 82 (2022), p. 102625.

3. RESEARCH PROBLEM

Colorectal cancer has one of the highest incidence rates around the world. Despite the colonoscopy being the standard diagnosis test, there are reports of missing polyp detection between 6-25%, especially at early stages. Computational approaches have emerged as a potential alternative to localize, segment, and characterize polyps, being a fundamental tool to support diagnosis and treatments. Currently, such strategies are based on deep representations, with remarked results but in bounded datasets, where only colonoscopy frames with polyps are considered. Such deep representations also remain limited to supporting polyp analysis in long sequences with non-controlled navigation conditions. Long colonoscopies may incorporate additional challenges, such as the incidence of light, specular reflections, intestinal folds, or water bubbles. In addition, most of the developed strategies are fully supervised, losing non-annotated colonoscopy frames that may be fundamental in discriminative frameworks to discard non-polyp regions. In fact, during colonoscopy procedures, the polyps only appear in short periods, with relatively few sizes concerning the intestinal tract considered as background, being a potential cause of false-positive detections.

Research question: How to detect and delineate polyps in long colonoscopy sequences from a restricted set of annotations?

4. OBJECTIVES

General Objective

- To propose a weakly supervised strategy to segment polyps in continuous colonoscopy sequences.

Specific Objectives

- To capture a dataset with continuous colonoscopy sequences captured during clinical procedures.
- To develop a deep representation to recover polyp segmentation from a trained set.
- To train deep representation into a weakly supervised scheme, approaching non-labeled information.
- To validate the proposed method in clinical scenarios using segmentation metrics.

5. PROPOSED APPROACH

This work introduces a weakly supervised scheme that takes advantage of background colonoscopy frames to improve polyp representation. Firstly, a backbone convolutional network is adapted to represent colonoscopy frames with and without polyps. The last convolutional layers are defined as the input to an attention module, implemented to focus on polyp masses, ignoring intestinal folds, specular reflections, and other regions present during colonoscopy procedures. Then, a sigmoid function and threshold over this branch result in the segmentation for each frame. In addition, a multitask loss function was proposed to consider polyps but also to monitor detection tasks and enhance boundaries of suspicious masses. The general pipeline of the proposed architecture is illustrated in Figure 7. The complete content of this section was accepted in the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) 2022[40].

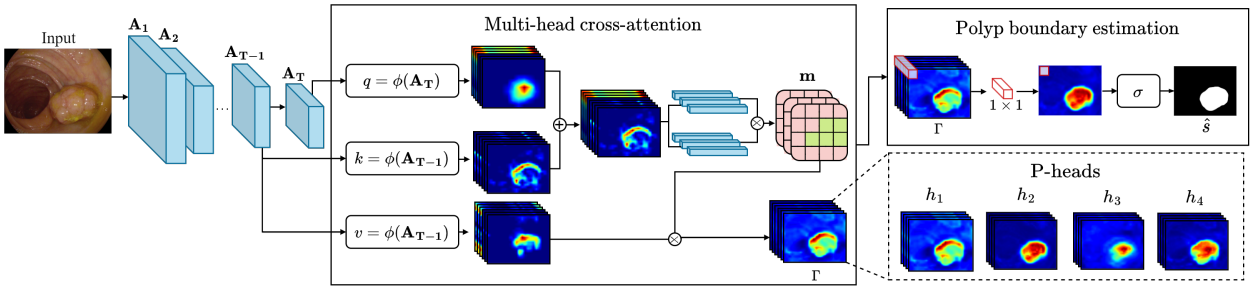


Figure 7. Proposed approach. With colonoscopy frames as input, a feature extraction process is performed from the last convolutional layers of the architecture. A set of dilated convolutions is applied to the extracted features, which are processed through attention modules.

---

[40] Lina RUIZ and Fabio MARTÍNEZ. "Weakly Supervised Polyp Segmentation from an Attention Receptive Field Mechanism". In: *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 2022, pp. 3745–3748.

5.1. Deep convolutional backbone representation

A primary representation of the intestinal tract, including polyps, is obtained using a deep convolutional backbone. These representations have demonstrated a robust capability to characterize complex local visual patterns with remarked variability among captured observations. The convolutional bank of features is robust for approximating polyp appearance and geometric description because these abnormal masses are well known for the reported strong variability and the non-parametric nature. Particularly, this work considers a colonoscopy sequence $\mathbf{I_t}(\mathbf{x})$ with size $\mathbf{x} \in \mathbb{R}^{W \times H}$, where each frame I_i is labeled with polyp ($y = 1$) or background ($y = 0$). Each colonoscopy frame is defined as an independent observation without temporal dependencies.

Hence, the convolutional architecture learns non-linear transformations that are progressively and hierarchically organized to compute a high-level projection of colonoscopy frame observations. First, the convolutional net is adjusted from a pre-trained architecture, followed by updating the representation from a weakly supervised scheme that includes labeled frames (polyp samples) and the intestinal tract as background. The two top layers are taken as polyp representations (see Figure 7). In such case, the convolutional activation outputs from these layers, defined as $\{\mathbf{A_{T-1}}, \mathbf{A_T}\}$ with $\mathbf{A_i} \in \mathbb{R}^{W' \times H' \times N_T}$, where $(W' \times H')$ is the spatial size with N_T activation maps for each activation block. It should be noted that the number of activations increases from the deeper direction of the architecture ($|N_T| > |N_{T-1}| > |N_{T-2}|$).

5.2. Multi-head attention from receptive field blocks

An attention-receptive field mechanism is proposed to project polyp convolution representation and robust representation from non-local modeling. The set of convolution activations $\mathbf{A_i}$ are mapped through RFB as $\phi(\mathbf{A_i})$, expressed with a set of dilated convolutions with different kernel sizes (κ), padding, and dilation. From these projections, responses

are obtained at different receptive field scales, learning, among others, larger neighborhoods, and multi-scale textural patterns.

Hence, three independent RFB branches are obtained from activations $\mathbf{A_{T-1}}$ and $\mathbf{A_T}$ at layers $T-1$ and T, respectively. Specifically, there is defined a query branch $q = \phi(\mathbf{A_T}) \in \mathbb{N}^{W'_T \times H'_T \times N_T}$ with a total of N_T activations and spatial size of $W'_T \times H'_T$. From layer $T-1$, two independent RFB branches are obtained, defined as the key $(k = \phi(\mathbf{A_{T-1}}))$ and value $(v = \phi(\mathbf{A_{T-1}})) \in \mathbb{N}^{W'_{T-1} \times H'_{T-1} \times N_{T-1}}$, where N_{T-1} represents the number of activations. Note that such projections preserve spatial information, and an up-sampling process is carried out to scale features of different layers. Then, the key and query branches are concatenated in a block $\theta(\phi(\mathbf{A_{T;T-1}}))$ with size $(W'_{T-1} \times H'_{T-1} \times N_c)$, where $N_c = |N_{T-1} + N_T|$. From this block, the attention matrix $(\mathbf{m})$ is calculated as a self-similarity between the pixels present in the projection of θ and θ^T. This projection is obtained from a dense attention matrix, constructed from the cosine distance, as:

$$\mathbf{m} = \frac{\theta(\phi(\mathbf{A_{T;T-1}}))^T \theta(\phi(\mathbf{A_{T;T-1}}))}{\|\theta(\phi(\mathbf{A_{T;T-1}}))\| \cdot \|\theta(\phi(\mathbf{A_{T;T-1}}))\|} \tag{2}$$

The value branch (v) is improved by following a matrix multiplication among the attention matrix $(\mathbf{m})$ and this branch, defined as the feature map $\Gamma = v \times \mathbf{m}$. Then, a set of 1×1 convolutions are applied to Γ over the channels to obtain a single map. The binary mask representing the polyp region is obtained after applying the sigmoid function (σ). Interestingly, this version stands out in polyp shape, allowing one to carry out a max-pooling operation with the sigmoid function to obtain the class to which the frame belongs $(\hat{y})$.

Then, multiple attention mechanisms in parallel are computed to enhance polyp representation with several maps dedicated to recovering polyp features. This multi-head attention define $\mathbf{Q} \in \mathbb{N}^{P \times D_T}, \mathbf{K} \in \mathbb{N}^{P \times D_{T-1}}, \mathbf{V} \in \mathbb{N}^{P \times D_{T-1}}$, where P represents the number of heads

extracted in the model, $D_T = \{W'_T \times H'_T \times N_T\}$ and $D_{T-1} = \{W'_{T-1} \times H'_{T-1} \times N_{T-1}\}$. Therefore, each head is defined as $head_p = Attention(q, k, v)$, which generates different sub-space representations. The multi-head attention output is the concatenation maps defined as $Multi - head(Q, K, V) = Concat(head_1, \cdots, head_p)$. This mechanism achieves non-local modeling of polyp observations, enhancing the morphological mass features. From such a mechanism, it is expected to refine the polyp edge and characterize shapes from the non-local representation that correlates all pixels into a particular frame.

5.3. A multitask polyp loss function

In this work, we analyze more realistic colonoscopy sequences with sparsely distributed polyps, which, in turn, have a huge shape variability. The deep representation should be capable of detecting polyps in independent frames and recovering the associated shape described from their contours. To adapt the strategy with such ability, each estimated polyp mask $(\hat{s})$ should overlap with the ground truth shape (s). Typically, to measure such matching, the Intersection over Union (IoU) is expressed as $IoU = \frac{\hat{s} \cap s}{\hat{s} \cup s}$. In our case, we reinforce the proposed network to recover shapes, following the IoU loss and considering the polyp boundaries. For doing so, the loss function for the polyp segmentation task is defined as $L_{seg} = L_{IoU} + L_{wIoU}$, where $L_{IoU} = 1 - IoU$ measure global overlapping, and the weighted IoU is formulated as:

$$L_{wIoU} = 1 - (1 + \gamma\alpha_{ij})(slog(\hat{s}) + IoU)^{1/2} \tag{3}$$

being γ a weighted hyperparameter, α calculates the difference between the central pixel (i, j) and its neighboring areas, and $slog(\hat{s})$ is a per-pixel binary cross-entropy[41].

---

[41] Jun WEI, Shuhui WANG, and Qingming HUANG. "F^{3}Net: fusion, feedback and focus for salient object detection". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 07. 2020,

Complementary, as an auxiliary task, from the segmentation, a global minimization rule is defined to classify according to frames with polyp $(y = 1)$ and background $(y = 0)$. The $\hat{y}$ is obtained from the Γ map using an adaptive max-pooling and a sigmoid function. In such a case, a binary cross-entropy is implemented between the polyp classification predicted ($\hat{y}$) and the actual class (y). This cross-entropy rule is then defined as $L_{cls} = ylog(\hat{y})$. For negative frames, without polyps $(y = 0)$, the function L_{seg} is fixed in one. From both rules, the loss function (ℓ) is defined as $\ell = L_{seg} + L_{cls}$.

pp. 12321–12328.

6. DATA AND EXPERIMENTAL SETUP

6.1. COLON dataset

A main contribution of this work is the capture and collection of the **COLON:** the largest **CO**lonoscopy **LON**g sequence dataset with more than 30 thousand labeled polyp images and 300 thousand intestinal fold background frames. Typical open databases only provide independent polyp images from short sequences (300 frames on average) focusing on polyp observations. Training and validation of the computational strategies over such restricted databases may limit the performance analysis in real scenarios. A routine colonoscopy easily exceeds a thousand frames to be processed, which reports more challenging visual variations. Therefore, these public datasets present polyp observations from isolated images or short sequences of several colonoscopy procedures. Despite reported variability, many strategies obtain polyp annotation scores up to 90%, which looks to overcome the problem of shape characterization. However, typical colonoscopies are very long and exhaustive procedures (taking more than 20 minutes per procedure), where the polyp observations appear in less than 10% of the frames. Beyond the textural and light variations, many of these procedures have additional challenges as poor patient preparation and intestinal folds that may be confused with polyp masses. So, far from resolving the polyp characterization problem, the community needs to develop new strategies to operate in scenarios closer to clinical procedures that consider many of the challenges during colonoscopies.

Therefore, the **COLON** dataset was generated with long colonoscopies collected from March 2021 to December 2022 in collaboration with the clinical center: *Instituto de Gastroenterología y Hepatología del Oriente* - IGHO S.A.S and the Biomedical Imaging, Vision, and Learning Laboratory - BIVL^{2}ab from *Universidad Industrial de Santander*. This

new dataset bridges the gap to validate approaches in more realistic sequences and close to the colonoscopy procedure. The study was carried out with 25 video sequences of independent colonoscopies, with videos without polyps (5 videos), with a polyp (18 videos), and with two polyps (2 videos). Each collected video sequence has an average of 16 thousand frames, with unbalanced observations, *i.e.,* most frames have no polyps. The colonoscopies were captured from an Olympus Evis Excera III 190 colonoscopy, with a spatial resolution of 720×480 and a temporal resolution 30 *fps*. Figure 8 shows examples of frames of different video sequences with the respective polyp delineations (green regions). These polyps have remarked morphological, size, and texture variability. In the last row, the blue arrows indicate some structures (specular reflections or mucous) with similar polyp patterns. The recorded videos are part of typical procedures, over uncontrolled scenarios, with some patients with poor preparation, several polyp-like structures, and strong camera movements.

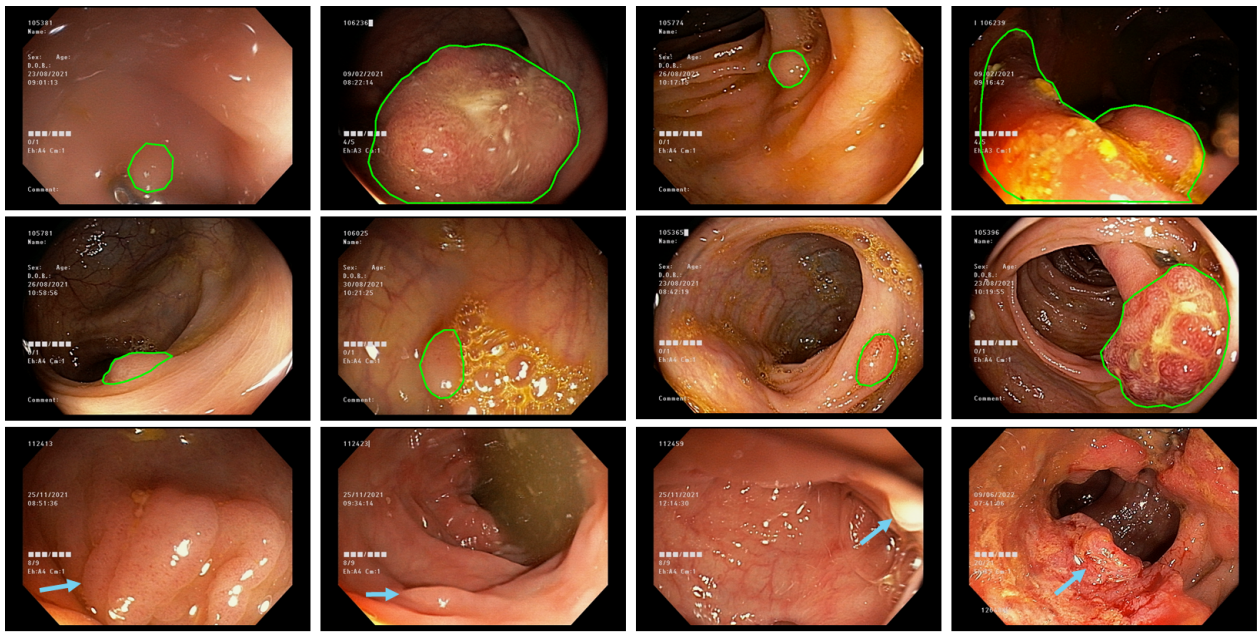


Figure 8. Frames extracted from the captured colonoscopy sequences. The first two rows contain polyps with their respective marking (green contour). The last row shows regions with similar polyp patterns.

To the best of our knowledge, the proposed dataset collects the major number of polyps and intestinal tract frames. Figure 9 compares the proposed dataset with baseline public datasets according to the number of frames. Classical datasets include isolated images without background information (Kvasir, ETIS-Larib, and others), while some datasets have collected short video sequences but focus mainly on polyp observations (ASU-Mayo [42] and CVC-Video [43]). The proposed dataset exceeds the amount of available data and labeled information, in addition to including the segmentation task in long sequences.

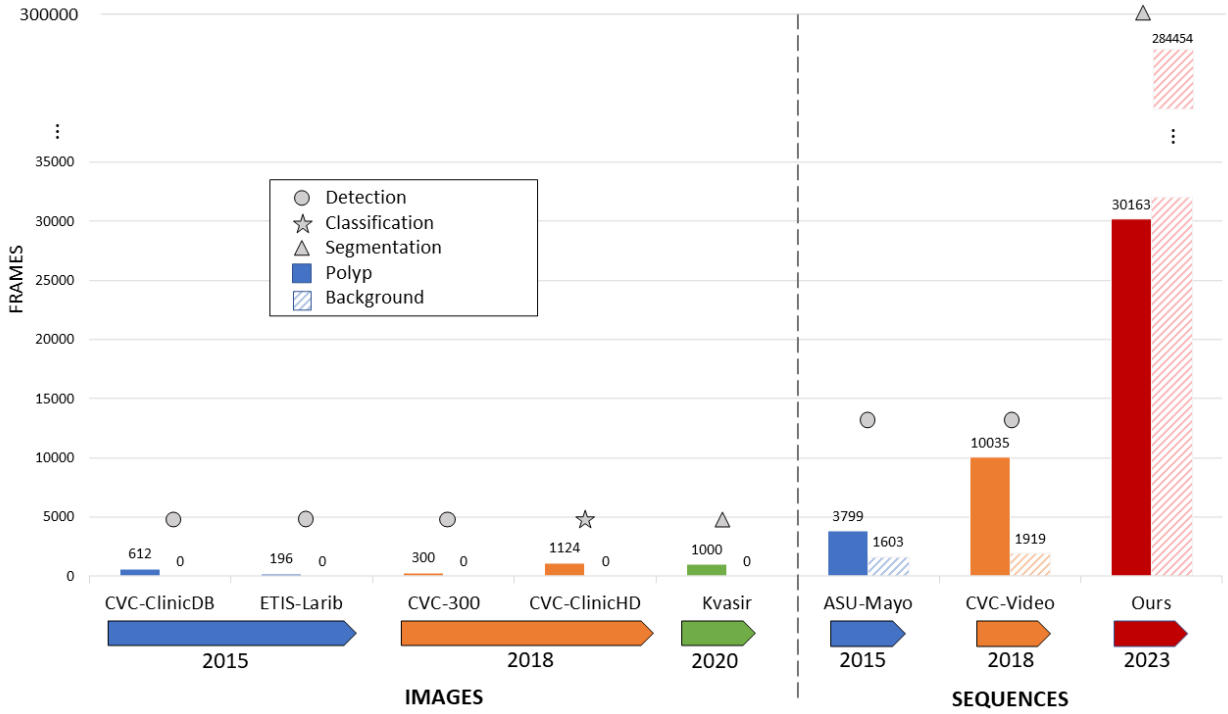


Figure 9. Comparison between public datasets (CVC-ClinicDB[44], ETIS-Larib[45], CVC-300[46], CVC-ClinicHD[47], Kvasir[48], ASU-Mayo[49], CVC-Video[50] available since 2015 and our proposed dataset for 2023.

[42] https://polyp.grand-challenge.org/AsuMayo/

[43] https://giana.grand-challenge.org/PolypDetection/

6.1.1. Morphological and pathological polyp features Each polyp was characterized according to size, from tiny (less than 10 mm) to large polyps (up to 40 mm), whose morphology can be sessile or pedunculated. Also, these masses were characterized according to NICE classification (Narrow-band images International Colorectal Endoscopic). This observational categorization focuses on vascular patterns, color, and lining surface being the type I hyperplastic, type II adenomas, and type III invasive carcinomas[51]. The pathology of some of these polyps was confirmed with a biopsy (adenoma, hyperplastic), showing the variability concerning observational analysis. As a demographic description, each colonoscopy associated with independent patients was also described according to gender and age ($[35-95], \mu = 65, \sigma = 13$). Figure 10 summarizes the morphological, pathological, and demographic information associated with the recorded procedures.

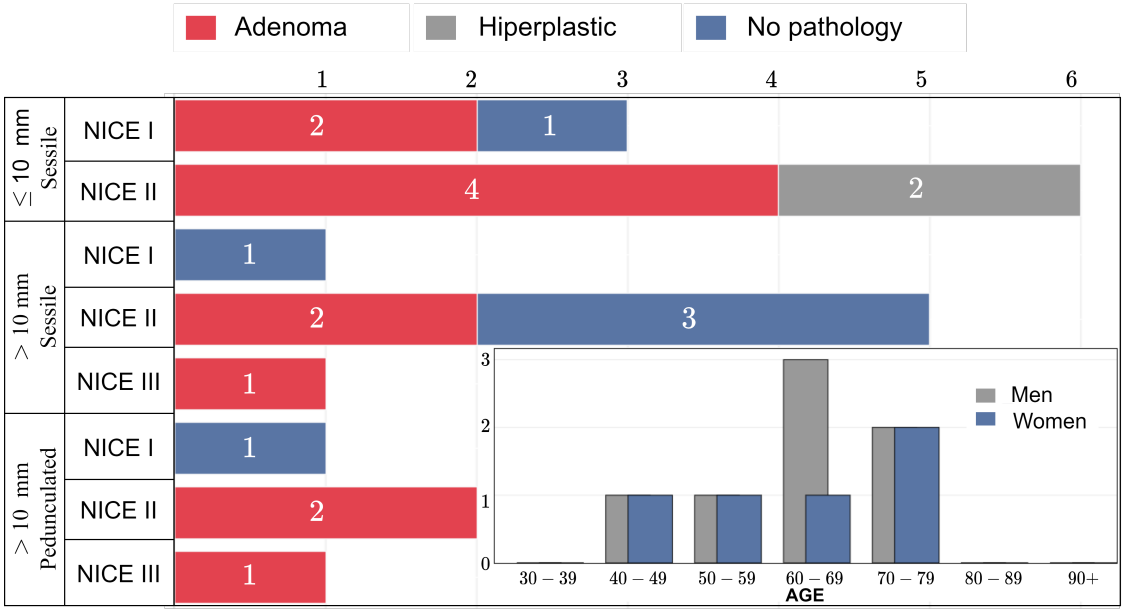


Figure 10. Polyp features distribution related to the size, morphology, NICE classification, and pathology result. At the bottom are the demographic patient details.

[51] Josipa PATRUN et al. "Diagnostic accuracy of NICE classification system for optical recognition of predictive morphology of colorectal polyps". In: *Gastroenterology research and practice* 2018 (2018).

6.2. External data

- **Kvasir-SEG:** This dataset was published in 2020. The Kvasir-SEG dataset contains 1000 polyp images and their corresponding segmentation masks extracted from different colonoscopy procedures[52].

- **ETIS-Larib:** This dataset was published in the Automatic Polyp Detection challenge in 2015. The ETIS-Larib dataset contains 196 frames with polyps and their corresponding ground truth extracted from 34 colonoscopy procedures[53].

- **CVC-Clinic DB:** This dataset was published in the Automatic Polyp Detection challenge in 2015. The CVC-Clinic dataset contains 612 polyp images with their corresponding binary mask extracted from 31 colonoscopy procedures[54].

- **ASU-Mayo Clinic Video:** This dataset was published in the Automatic Polyp Detection challenge in 2015. The ASU-Mayo dataset contains 20 training videos with the corresponding polyp mask divided into 10 short sequences of colonoscopy with polyps and 10 videos with only the intestinal tract. Each video lasts one minute on average. Frames that do not present polyps have as a ground truth a black mask. In addition, this dataset includes 18 testing videos without ground truth annotations[55].

---

[52] Konstantin POGORELOV et al. "KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection". In: *Proceedings of the 8th ACM on Multimedia Systems Conference*. 2017, pp. 164–169.

[53] Juan SILVA et al. "Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer". In: *International Journal of Computer Assisted Radiology and Surgery* 9.2 (2014), pp. 283–293.

[54] Jorge BERNAL et al. "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians". In: *Computerized medical imaging and graphics* 43 (2015), pp. 99–111.

[55] Nima TAJBAKHSH, Suryakanth R. GURUDU, and Jianming LIANG. "Automated polyp detection in colonoscopy videos using shape and context information". In: *IEEE Transactions on Medical Imaging*

- **CVC-Video Clinic DB:** This dataset was published in the Endoscopic Vision Challenge in 2018. The CVC-Video dataset contains 18 short colonoscopy sequences extracted from different procedures, which last a minute on average. All videos present a polyp inside and the corresponding ground truth. The frames that do not present polyps have as a ground truth a black mask. Moreover, this dataset includes 18 testing videos without ground truth annotations[56].

6.3. Validation

The proposed method was exhaustively validated from different public datasets but also regarding the owner dataset that considers relatively long colonoscopy sequences. Figure 11 represented the general training and validation scheme. Firstly, we use training splits of a total of three public datasets: 1) polyp frames (1450) from Kvasir-SEG and CVC-Clinic; 2) background frames (1450 frames taken from 10 sequences without polyps) from ASU-Mayo. Then, fine-tuning is performed from the owner training split using 20 long sequences with polyps (labeled every ten frames) and 5 background videos, extracting 1450 frames of each considered class.

For the test, we consolidated validation from five public datasets: Kvasir-SEG, ETIS-Larib, CVC-Clinic, ASU-Mayo, and CVC-Video. Secondly, the proposed strategy was validated concerning the introduced long sequence setup. In such a case, we consider a total of 10 videos with polyp annotations labeled each frame that contains this region ($>$ 28 thousand labels) and 5 background videos ($>$ 48 thousand frames). Additionally, we include a baseline comparison with two state-of-the-art architectures dedicated to polyp segmentation

---

35.2 (2015), pp. 630–644.

[56] Quentin ANGERMANN et al. “Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis”. In: *Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures*. Springer. 2017, pp. 29–41.

(PraNet[57] and HarDNet[58]).

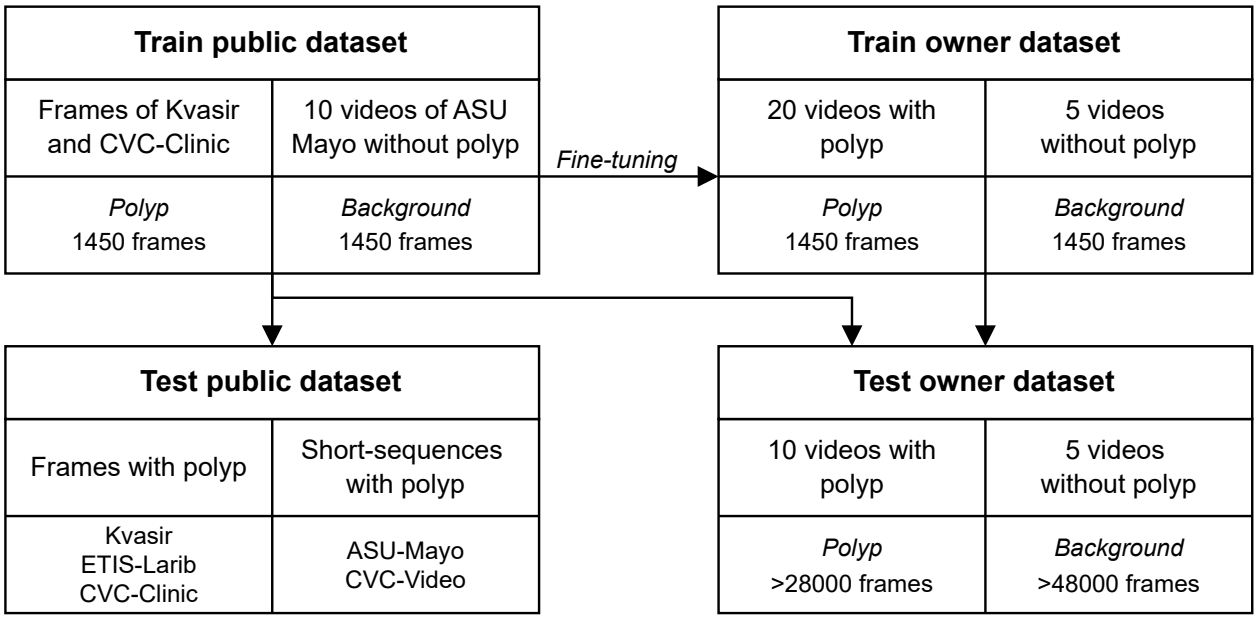


Figure 11. Dataset distribution for training and testing the model.

6.4. Network configuration

The proposed approach was validated according to the following configuration. Regarding the convolutional backbone, the Res2Net-50 was adopted to decompose the local information of input polyp frames[59]. Such a network includes multi-scale features through hierarchical residual-like connections, allowing a significant information decomposition that may be essential to discriminating close textural patterns among intestinal folds and polyps. Concerning the multiple head attentions, the proposal was validated with heads

---

[57] FAN et al., “Pranet: Parallel reverse attention network for polyp segmentation”.

[58] HUANG, WU, and LIN, “Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps”.

[59] Shang-Hua GAO et al. “Res2net: A new multi-scale backbone architecture”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.2 (2019), pp. 652–662.

$P = \{3, 4, 5\}$ and including RFB modules with dilated convolutions, whose kernel sizes were defined as $\kappa = \{1, 3, 5, 7\}$.

For training, we implemented a weakly supervised scheme taking intestinal tract frames without polyp labels. For each experiment, was set a total of 1450 polyp images and 1450 without observed polyps. The training and testing size for all frames in the selected datasets was 352×352. A curriculum learning regime starts with a $1e^{-4}$ learning rate and 80 epochs for the first training. Then, the fine-tuning strategy was adjusted with a learning rate of $1e^{-5}$ and 30 epochs. For both architectures, the optimizer was established as the Adam algorithm. Statistical differences between our proposed approach and the baseline methods were calculated using the Mann-Whitney test.

7. EVALUATION AND RESULTS

Firstly, an ablation study was carried out concerning the contribution of multiple cross-attention $P-$heads ($P = \{3, 4, 5\}$). This validation was performed over two public datasets of short colonoscopy sequences (CVC-Video and ASU-Mayo). Table 1 summarizes the contribution of extracting several head attention mechanisms. As observed, the recovery at each head contributes to better polyp segmentation. When four heads are implemented, the best precision record is achieved, up to 90%. Contrarily, for five heads in the ASU-Mayo dataset, a low recall and IoU are reported (28% and 26%, respectively). Further experiments were set with four heads.

Table 1. Ablation study of the multi-head cross-attention mechanism over short sequences datasets. The number of $P-$heads varies in each experiment.

Heads	**Dataset**	**Metrics(%)**			
		Recall	**Precision**	**Specificity**	**IoU**
P=3	CVC-Video	60.7	87.7	63.8	47.0
	ASU-Mayo	56.9	74.3	51.2	47.3
P=4	CVC-Video	**61.9**	**96.4**	**90.1**	**47.3**
	ASU-Mayo	**60.9**	92.4	87.6	**50.6**
P=5	CVC-Video	60.5	94.5	84.9	47.1
	ASU-Mayo	28.2	**98.1**	**98.6**	26.1

The proposed approach was firstly validated w.r.t the capability to recover segmentation from isolated colonoscopy frames. In such a case, the Kvasir-SEG dataset was used as a reference to compare with other state-of-the-art strategies. Table 2 reports the performance achieved by the proposed method and different baseline methodologies. In summary, the proposed approach achieves competitive results, outperforming several strate-

gies (ColonSegNet[60] and FANet[61] at around 9% for recall and 14% for IoU). On the other hand, the HarDNet achieves the best performance, probably due to the specialization to always identify polyps in all frames.

Table 2. Comparison with respect to the state-of-the-art methods using the Kvasir-SEG dataset.

Architecture	**Recall (%)**	**IoU (%)**
PraNet[62]	92.0	83.7
HarDNet[63]	**95.0**	**86.2**
ColonSegNet[64]	81.9	69.8
FANet[65]	85.0	69.7
CaraNet[66]	92.9	80.0
DilatedSegNet[67]	91.7	83.4
Proposed approach	93.0	83.6

Secondly, the proposed approach was validated in short colonoscopy sequences. In such cases, we extended the validation to include precision and specificity for measuring the performance of the models against frames with and without polyps. This baseline comparison was carried out over the CVC-Video Database. Table 3 summarizes results obtained for the proposed approach in relation to the state-of-the-art. The proposed strategy achieves a remarkable performance in precision and recall (96.4% and 90.1%, respectively), evidencing capabilities to characterize polyp morphology and obtain a lower amount of false positives.

To validate the generalization capability of the proposed approach, we extend validation for additional three datasets. For training, we use Kvasir-SEG, CVC-Clinic, and ASU-Mayo datasets. Then, we test with the following unseen datasets: ETIS-Larib and CVC-

---

[60] JHA et al., "Real-time polyp detection, localization and segmentation in colonoscopy using deep learning".

[61] Nikhil Kumar TOMAR et al. "Fanet: A feedback attention network for improved biomedical image segmentation". In: *IEEE Transactions on Neural Networks and Learning Systems* (2022).

Table 3. Comparison with respect to the state-of-the-art methods using the CVC-Video Colon Database (CVC-Video).

Architecture	Metrics(%)			
	Recall	**Precision**	**Specificity**	**IoU**
PraNet[68]	60.8	81.4	40.9	48.4
HarDNet[69]	**66.2**	87.3	59.0	51.6
ResUNet++, TTA[70]	21.1	33.4	-	64.4
ResUNet++, TTA, CRF[71]	20.7	30.4	-	**64.7**
ResUNet++, CRF[72]	22.3	30.6	-	64.1
Proposed approach	61.9	**96.4**	**90.1**	47.4

Video. In such a case, we include architectures (PraNet, HarDNet) trained under the same scheme as the proposed approach. Table 4 summarizes the achieved results for the three considered architectures and the five public datasets. Regarding the CVC-Clinic dataset, the proposed approach evidence a proper polyp segmentation (recall of 94.2% and IoU of 79.9%), reporting non-statistical differences for PraNet ($p < 0.6$) and HarDNet ($p < 0.2$) architectures.

Polyp characterization is a principal task for early diagnosis, and computational support may help to reduce false negative scores (around 6-25%, as reported in the literature[73]). Therefore, the proposed approach was validated with short sequence datasets for polyp localization in scenarios with background frames (ASU-Mayo and CVC-Video). The proposed method outperforms the specificity with a gain of around 31% (CVC-Video) and 20% (ASU-Mayo) regarding the HarDNet architecture. These metrics count a remarkable set of intestinal folds marked as polyps, which may produce a significant latency in the analysis of colonoscopies. Concerning the PraNet, the proposed strategy achieves a specificity gain of 42% (ASU-Mayo) and 50% (CVC-Video). Interestingly, the preci-

[73] ANGERMANN, HISTACE, and ROMAIN, “Active learning for real time detection of polyps in video-colonoscopy”.

sion of the proposed approach has an average of 12% (CVC-Video) over the baseline architectures, which induces enhancing the capability approach to recover correct polyps concerning false polyps delineated by the computational methods. The outperforming of the proposed method may be associated with the modeling of weakly supervised learning, which introduces a regulator into minimization cost to include background frames without polyp information. Following the statistical test, a significant difference was found in the achieved results for the proposed approach with respect to the baselines and the CVC-Video dataset. Considering the precision ($p < 0.05$) and specificity ($p < 0.01$) for PraNet architecture, while for the HarDNet method, the statistical differences for precision ($p < 0.01$) and specificity ($p < 1e^{-4}$) were more significant.

Table 4. State-of-the-art comparison in polyp segmentation over five public datasets. Statistical differences between the baselines and our proposed approach were measured. Non-statistical differences are denoted by †, while significant statistical variations by *.

Method	**Dataset**	**Metrics (%)**			
		Recall	**Precision**	**Specificity**	**IoU**
PraNet[74]	Kvasir-SEG	92.0	-	-	83.7
	CVC-Clinic	91.8	-	-	82.3†
	ETIS-Larib	66.8	-	-	58.4
	ASU-Mayo	64.1	74.6†	45.8*	55.9
	CVC-Video	60.8	81.4*	40.9*	48.4
HarDNet[75]	Kvasir-SEG	**95.0**	-	-	**86.2**
	CVC-Clinic	94.2	-	-	**85.5**†
	ETIS-Larib	**71.0**	-	-	**61.5**
	ASU-Mayo	**66.7**	83.4†	67.2*	**56**
	CVC-Video	**66.2**	87.3*	59.0*	**51.6**
Proposed approach	Kvasir-SEG	93.0	-	-	83.6
	CVC-Clinic	**94.2**	-	-	79.9
	ETIS-Larib	59.2	-	-	52.4
	ASU-Mayo	60.9	**92.4**	**87.6**	50.6
	CVC-Video	61.9	**96.4**	**90.1**	47.3

7.1. Evaluation from COLON: the long colonoscopy videos dataset

The main interest in this study was to validate computational strategies in scenarios closer to standard colonoscopy environments, which include sudden movements, poor patient preparation, and much of the observed frames corresponding to background information. Then, the proposed approach and the baseline strategies were run over long videos captured in the owner dataset. All methods considered were trained with the public dataset and validated over the test owner dataset (as depicted in Figure 11).

A deep analysis of achieved results was obtained considering the performance at different polyps groups, reached according to morphological variables such as the size, morphology, NICE, and pathology stratification. Figure 12 summarizes segmentation results from violin plots according to each considered morphological group. As expected, regarding morphological characterization, the proposed approach shows outstanding performance in big polyps ($>$ 10 mm) with sessile features. Also, a remarkable average IoU of 62% for tiny polyps was achieved, as well as an average of 67% for pedunculated ones. The statistical test shows differences among validated sub-groups suggesting a remarkable trend of the proposed approach to delineate larger polyps with marked texture. Other sub-types remain challenging because polyps remain adhered to the intestinal tract.

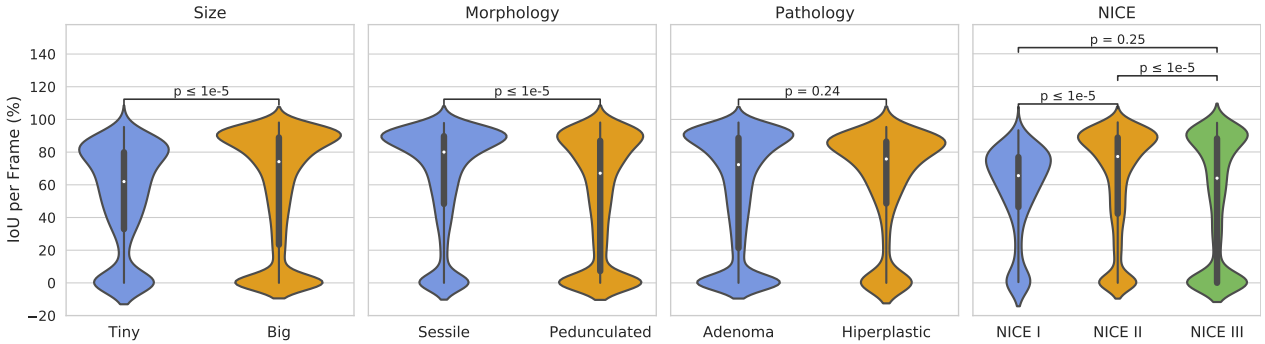


Figure 12. Evaluation of the proposed method on the owner dataset. Each column presents the IoU obtained according to the data categorization (size, morphology, pathology, and NICE).

Interestingly, the results of segmentation regarding adenoma ($\mu = 72 \pm 36$) and hiperplas-

tic ($\mu = 75 \pm 32$) classes were statistically equivalent ($p = 0.24$). Despite of vascularity of such polyps may have remarked differences, the proposed approach achieves a stable representation. In the same line, regarding NICE classification, the proposed approach retrieves the median IoU of NICE I ($\mu = 65 \pm 27$), NICE II ($\mu = 77 \pm 33$), and NICE III ($\mu = 64 \pm 38$). It should be noted that such classification is only achieved with NBI light, and there was no statistical significance among sub-types of classification. An important remark in this validation is the evidence of significantly lower IoU obtained in more challenging scenarios suggesting a remaining gap between computational approximation and real colonoscopy scenarios.

A further analysis was carried out by comparing the performance of the proposed strategy concerning the PraNet and HarDNet architectures and considering size sub-groups because of the clinical relevance. Figure 13 shows the results achieved by the three strategies. The upper plot illustrates the IoU distribution obtained for the methods according to polyp size. The horizontal red line corresponds to the median value of the proposed approach. Remarkably, the proposed method performed better in recovering the segmentation of polyps with sizes of less than 10 mm. Contrarily, the PraNet has a significant set of samples (30% of IoU) with lower scores or without polyp detection. Complementary, in the second row of Figure 13 is summarized the precision, recall, and specificity for each method included in this analysis. The high data imbalance to emulate real scenarios is remarkable, where a considerable part of the frames correspond to the intestinal background without evidence of polyps.

An additional experiment was carried out by adjusting the representation from the training set of the owner dataset. For doing so, the architectures were adjusted from a fine-tuning process from complementary 80 epochs but using a learning rate of $1e^{-6}$. Table 5 illustrates the achieved results for the proposed approach and baseline architectures. Table 5-top summarizes the result achieved of architectures using external datasets for

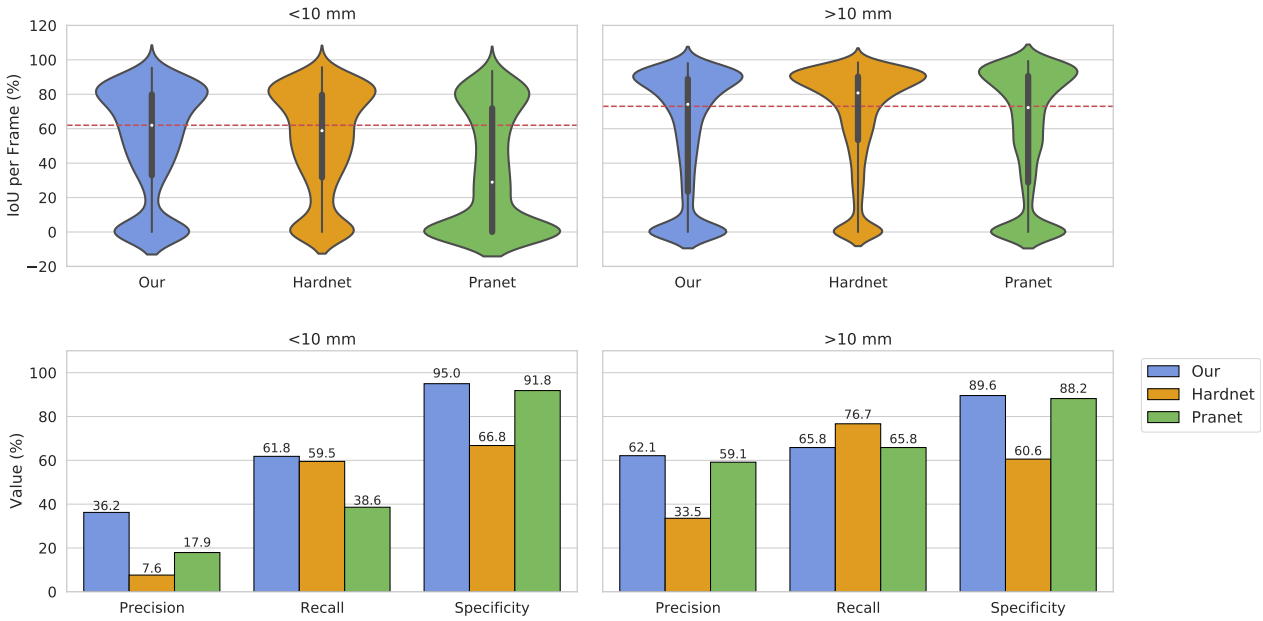


Figure 13. Comparison of the state-of-the-art strategies with our method using the proposed dataset to validate. The first row presents the IoU per frame for each method, while the second row shows the precision, recall, and specificity metrics.

training and evidencing the generalization capability of each method. The proposed approach outperforms the other architectures in precision and specificity, showing the ability to adopt background information and exclude false positive detections. Even though HarDNet presents a better IoU and recall, there is a clear performance imbalance regarding precision. There is a significant set of false positive detection, which in clinical practice may induce a considerable latency of the approach. Table 5-bottom is evidenced by the performance achieved for architectures once fine-tuning is carried out from training data of long video samples. Interestingly, the proposed approach achieves a dramatic precision gain of 20%, while the other architectures decrease the performance, a fact that may be associated with limitations to modeling the background. Overall, the proposed method achieved higher performance of such updating process, achieving coherent polyp detections and segmentations. For this experiment, the PraNet reports a better recall but affecting the precision, *i.e.,* a lower capability to detect polyps during the procedure.

Table 5. Analysis of fine-tuning training and model generalization using the test set of the owner dataset.

Train	Architecture	Metrics(%)			
		Recall	**Precision**	**Specificity**	**IoU**
Public datasets	Our	65.5	**58.5**	**91.6**	57.7
	PraNet	63.3	52.3	89.5	56.8
	HarDNet	**75.1**	26.8	62.8	**65.6**
Owner dataset	Our	75.5	**70.8**	**93.4**	65.2
	PraNet	**84.3**	21.2	34.0	**74.2**
	HarDNet	81.6	21.3	36.4	71.1

Qualitative analysis is shown in Figure 14, reporting the corresponding frame observation (in green is delineated the ground truth) with the attention map and the recovered binary mask. Despite the high variability in the shape of the polyps, the specular reflections, and the multiple vascular regions, the proposed approach achieves remarkable adaptability to retrieve correct polyp segmentation (see in Figure 14 first and second row). Also, the proposed method recognizes background frames, generating uniform attention maps without importance on a particular region, avoiding the detection of false positives (third row). The last example (fourth row) is illustrated a mistake sample regarding polyp segmentation. Although the attention map focuses on a significant polyp region, there exist limitations to properly delineating the shape because of the negligible boundaries regarding the intestinal tract and the high similarity these regions have in texture.

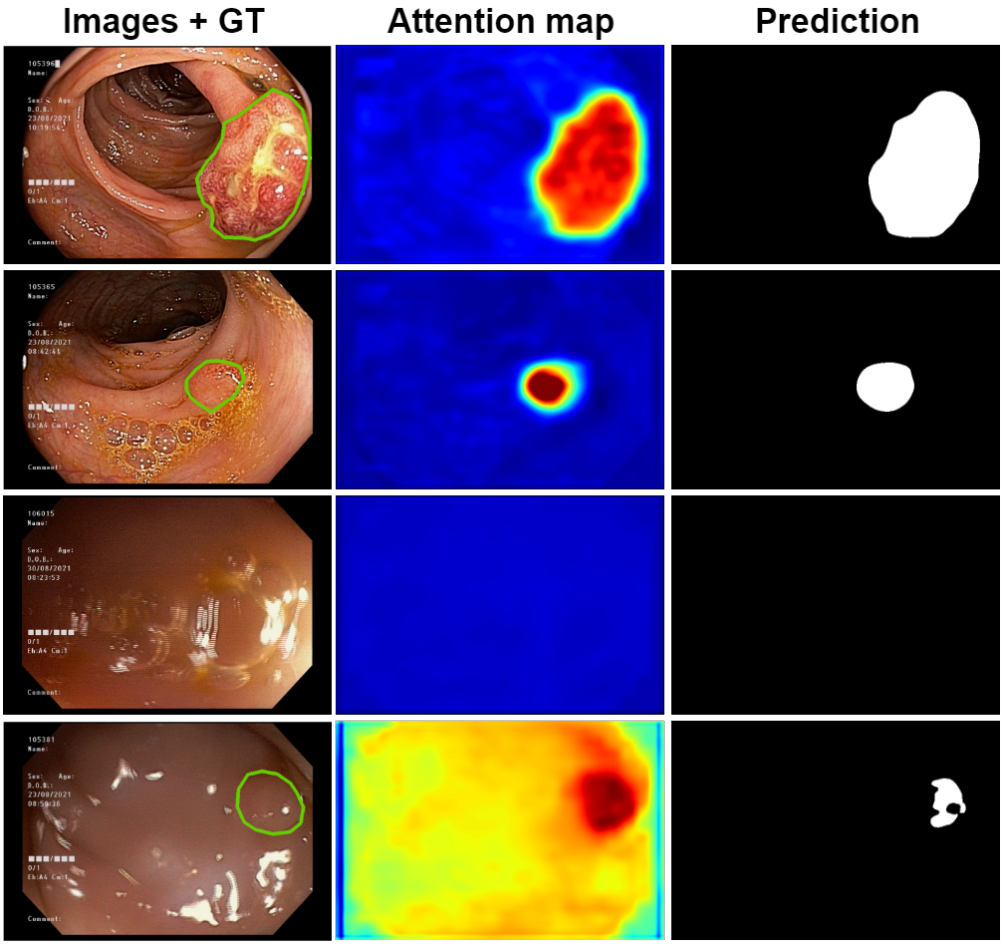


Figure 14. Qualitative results. The first and second rows show a polyp whose segmentation was appropriate. In the third row, an intestinal tract frame whose output was a black mask, and in the last row, a polyp that the method did not segment. a) Images with the ground truth (green delineation). b) Heatmap. c) Binary segmentation.

8. DISCUSSION

This work proposed a weakly supervised strategy that includes polyp frame delineations but also took advantage of fold intestinal frames without polyps (more than 90% of information in a colonoscopy). Under such a training scheme, a deep multi-head attention representation was adjusted to detect and segment polyps in long colonoscopy sequences. From an owner dataset with 10 long colonoscopies (on average 15.000 frames per video), the proposed strategy reaches a specificity of 93% and a precision of 70%, overcoming typical false-polyp detections and facing challenges related to intestinal folds, water bubbles, and specular reflections. Besides, regarding polyp segmentation, the proposed strategy achieved a recall of 75% and IoU of 65%.

Proposals in state-of-the-art have been principally addressed to segment polyps over independent frames or short sequences with almost always polyp presence. For instance, the HarDNet[76] and PraNet[77] architectures evidenced remarkable IoU scores, on average 77% and 74%, respectively, for the three polyp frames datasets. Nonetheless, the evaluation over long colonoscopies evidenced limitations of these architectures, such as low precision for tiny polyps (HarDNet = 7%; PraNet = 17%), despite the achieved score in per-frame and short sequences datasets. These results evidenced the limitation of approaches to be extended in clinical scenarios, especially over tiny polyps that are relevant for early cancer diagnosis. Contrarily, the proposed method achieved a more robust precision (36%), recall (61%), and specificity (95%) in such tiny polyps. Although HarDNet evidenced a significant gain regarding big polyps (larger than 10 mm), the specificity re-

---

[76] HUANG, WU, and LIN, “Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps”.

[77] FAN et al., “Pranet: Parallel reverse attention network for polyp segmentation”.

mains with a low performance, which suggests a difficulty to differentiate between frames with polyps and background. Furthermore, the proposed approach evidences a coherent segmentation for this type of polyps, reporting an average IoU of 75%.

It should be noted that the proposed approach and baseline-considered strategies were trained with external public datasets. These results evidence key advantages to cover a higher spectrum of colonoscopy observations, allowing for more generalization capabilities. Likewise, these strategies were fine-tuned with long colonoscopy videos, again showing a better adaptation for the proposed approach, increasing the precision to 12% and specificity to 2%. Such fact results are remarkable to support early detection from positive detections in long colonoscopy sequences. Locally, the polyp shape recovery was slightly better for HarDNet, reporting 9% more in IoU regarding the proposed approach.

Additional validation was performed using different public datasets at the frame and short sequence levels. Concerning the datasets with only polyp frames (Kvasir-SEG, ETIS-Larib, CVC-Clinic), the proposed approach achieved competitive results without significant statistical differences regarding the PraNet and HarDNet. Other methods such as FANet[78], CaraNet[79], and DilatedSegNet[80] were validated on the Kvasir dataset reporting an IoU of 69.7%, 80%, and 83.4%, respectively. In this case, the proposed method obtained better results than these approaches (83.6%). Concerning short sequence datasets (ASU-Mayo and CVC-Video), the proposed strategy reported a more consistent precision and specificity. Specifically, over the CVC-Video dataset, we obtain 50% more than PraNet in speci-

---

[78] TOMAR et al., “Fanet: A feedback attention network for improved biomedical image segmentation”.

[79] Ange LOU et al. “CaraNet: context axial reverse attention network for segmentation of small medical objects”. In: *Medical Imaging 2022: Image Processing*. Vol. 12032. SPIE. 2022, pp. 81–92.

[80] Nikhil Kumar TOMAR, Debesh JHA, and Ulas BAGCI. “DilatedSegNet: A Deep Dilated Segmentation Network for Polyp Segmentation”. In: *International Conference on Multimedia Modeling* (2023), pp. 334–344.

ficity, while 30% more than HarDNet. Therefore, the background frames included in the training allow a decrease in the false positive amount and, in addition, a better recognition between the intestinal tract and the polyp. Other approaches have also been validated in CVC-Video, achieving 64% in IoU. However, they present a lower precision (30%) and specificity (20%) with different configurations of ResUnet++[81]. These approaches are dedicated to segmentation tasks in frames with polyps, and specificity is not reported. Hence, the proposed method surpasses the baseline strategies showing competitive results over five public datasets, mainly in precision and specificity metrics. For instance, the proposed method obtained a precision of 96% and specificity of 90% for the CVC-Video, while for the ASU-Mayo dataset, the achieved results were 92% and 87%, respectively. In a comprehensive validation, the proposed approach demonstrated robust polyp segmentation and localization over long sequences in scenarios closer to clinical routine.

[81] JHA et al., “A Comprehensive Study on Colorectal Polyp Segmentation With ResUNet++, Conditional Random Field and Test-Time Augmentation”.

9. CONCLUSIONS AND FUTURE WORK

This work presented a weakly supervised strategy that approximates polyp segmentation on long colonoscopy sequences, learning not only from abnormal masses but also from the intestinal background. The introduced architecture implements a multi-head cross-attention module that allows coding polyp features from a multi-scale deep representation following local (from convolutions), regional (from RBFs), and non-local (from attention maps) patterns. Moreover, the minimization rule learned to differentiate between polyp images and background frames from a global and local perspective. In general, the learning of background features showed a low false positive amount in short and long sequences, outperforming the state-of-the-art in precision and specificity metrics. The proposed approach presented the best performance in recognizing polyps smaller than 10 mm in long colonoscopy sequences, where similarities with the intestinal tract are more evident.

Future work includes exploring new training schemes without further requirements of label annotations to avoid expert-subjectivity learning. In addition, designing a new computational mechanism to detect small masses over complete colonoscopies will be implemented and validated. Also, we expect to increase the variability and challenge of the dataset by incorporating more colonoscopies with additional polyp conditions (textural features and multiple polyps in observations, among others). In such a case, we expect to bring to the community a dataset that approximates approaches to clinical routine but also introduces new challenges to the polyp characterization area.

BIBLIOGRAPHY

ANGERMANN, Quentin et al. “Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis”. In: *Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures*. Springer. 2017, pp. 29–41.

ANGERMANN, Quentin, Aymeric HISTACE, and Olivier ROMAIN. “Active learning for real time detection of polyps in videocolonoscopy”. In: *Procedia Computer Science* 90 (2016), pp. 182–187.

BAHDANAU, Dzmitry, Kyunghyun CHO, and Yoshua BENGIO. “Neural machine translation by jointly learning to align and translate”. In: 2014.

BERNAL, Jorge et al. “GTCreator: a flexible annotation tool for image-based datasets”. In: *International Journal of Computer Assisted Radiology and Surgery (IJCARS)* 14.2 (2019), pp. 191–201.

– “Polyp segmentation method in colonoscopy videos by means of MSA-DOVA energy maps calculation”. In: *Clinical Image-Based Procedures. Translational Research in Medical Imaging*. Springer. 2014, pp. 41–49.

– “WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians”. In: *Computerized medical imaging and graphics* 43 (2015), pp. 99–111.

BERNAL, Jorge, Javier SÁNCHEZ, and Fernando VILARINO. “Towards automatic polyp detection with a polyp appearance model”. In: *Pattern Recognition* 45.9 (2012), pp. 3166–3182.

BRANDAO, Patrick. et al. “Fully convolutional neural networks for polyp segmentation in colonoscopy”. In: *Medical Imaging 2017: Computer-Aided Diagnosis*. Vol. 10134. SPIE. 2017, pp. 101–107.

CHEN, Long et al. “Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 5659–5667.

CHEN, Siwei, Gregor URBAN, and Pierre BALDI. “Weakly Supervised Polyp Segmentation in Colonoscopy Images Using Deep Neural Networks”. In: *Journal of Imaging* 8.5 (2022). ISSN: 2313-433X.

DONG, Bo et al. “Polyp-PVT: Polyp Segmentation with Pyramid Vision Transformers”. In: *arXiv e-prints, arXiv-2108* (2021).

FAN, Deng-Ping et al. “Pranet: Parallel reverse attention network for polyp segmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 263–273.

FERLAY, Jacques et al. “Cancer statistics for the year 2020: An overview”. In: *International Journal of Cancer* 149.4 (2021), pp. 778–789.

GAO, Haoqi and Koichi OGAWARA. “Adaptive data generation and bidirectional mapping for polyp images”. In: *IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*. IEEE. 2020, pp. 1–6.

GAO, Shang-Hua et al. “Res2net: A new multi-scale backbone architecture”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.2 (2019), pp. 652–662.

GREENBERGER, Norton et al. *Current Diagnosis and Treatment: Gastroenterology, Hepatology, and Endoscopy*. 3rd ed. Mc Graw Hill, 2016.

GUO, Xiaoqing et al. “Non-equivalent images and pixels: Confidence-aware resampling with meta-learning mixup for polyp segmentation”. In: *Medical Image Analysis* 78 (2022), p. 102394.

GUO, Yunbo, Jorge BERNAL, and Bogdan J. MATUSZEWSKI. “Polyp Segmentation with Fully Convolutional Deep Neural Networks Extended Evaluation Study”. In: *Journal of Imaging* 6.7 (2020).

HUANG, Chien-Hsiang, Hung-Yu WU, and Youn-Long LIN. “Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps”. In: *arXiv preprint arXiv:2101.07172* (2021).

HWANG, Sae et al. “Polyp Detection in Colonoscopy Video using Elliptical Shape Feature”. In: *2007 IEEE International Conference on Image Processing*. Vol. 2. 2007, pp. II–465.

JHA, Debesh et al. “A Comprehensive Study on Colorectal Polyp Segmentation With ResUNet++, Conditional Random Field and Test-Time Augmentation”. In: *IEEE Journal of Biomedical and Health Informatics* 25.6 (2021), pp. 2029–2040.

– “Real-time polyp detection, localization and segmentation in colonoscopy using deep learning”. In: *IEEE Access* 9 (2021), pp. 40496–40510.

JIA, Xiao et al. “Automatic Polyp Recognition in Colonoscopy Images Using Deep Learning and Two-Stage Pyramidal Feature Prediction”. In: *IEEE Transactions on Automation Science and Engineering* 17.3 (2020), pp. 1570–1584.

KANG, Jaeyong and Jeonghwan GWAK. “Ensemble of Instance Segmentation Models for Polyp Segmentation in Colonoscopy Images”. In: *IEEE Access* 7 (2019), pp. 26440–26447.

LAMBERT, R. “The Paris endoscopic classification of superficial neoplastic lesions: esophagus, stomach, and colon: November 30 to December 1, 2002”. In: *Gastrointestinal Endoscopy* 58 (2003), S3–S43.

LI, Qiaoliang et al. “Colorectal polyp segmentation using a fully convolutional neural network”. In: *2017 10th International Congress on Image and Signal Processing, Biomedical Engineering and Informatics (CISP-BMEI)*. IEEE. 2017, pp. 1–5.

LIN, Di et al. “Scribblesup: Scribble-supervised convolutional networks for semantic segmentation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 3159–3167.

LIU, Liangliang et al. “A survey on U-shaped networks in medical image segmentations”. In: *Neurocomputing* 409 (2020), pp. 244–258.

LIU, Songtao and Di HUANG. “Receptive field block net for accurate and fast object detection”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 385–400.

LONG, Jonathan, Evan SHELHAMER, and Trevor DARRELL. “Fully convolutional networks for semantic segmentation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 3431–3440.

LOU, Ange et al. “CaraNet: context axial reverse attention network for segmentation of small medical objects”. In: *Medical Imaging 2022: Image Processing*. Vol. 12032. SPIE. 2022, pp. 81–92.

ORTEGA-MORÁN, Juan et al. “Medical needs related to the endoscopic technology and colonoscopy for colorectal cancer diagnosis”. In: *BMC cancer* 21.1 (2021), pp. 1–12.

PARK, Kyeong-Beom and Jae Yeol LEE. “SwinE-Net: Hybrid deep learning approach to novel polyp segmentation using convolutional neural network and Swin Transformer”. In: *Journal of Computational Design and Engineering* 9.2 (Apr. 2022), pp. 616–632. ISSN: 2288-5048.

PATRUN, Josipa et al. “Diagnostic accuracy of NICE classification system for optical recognition of predictive morphology of colorectal polyps”. In: *Gastroenterology research and practice* 2018 (2018).

PÉREZ, Eduardo. *Gastroenterología*. 1st ed. McGraw Hill Mexico, 2012.

POGORELOV, Konstantin et al. “KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection”. In: *Proceedings of the 8th ACM on Multimedia Systems Conference*. 2017, pp. 164–169.

PUYAL, Juana González-Bueno et al. “Polyp detection on video colonoscopy using a hybrid 2d/3d cnn”. In: *Medical Image Analysis* 82 (2022), p. 102625.

QADIR, Hemin Ali et al. “Toward real-time polyp detection using fully CNNs for 2D Gaussian shapes prediction”. In: *Medical Image Analysis* 68 (2021), p. 101897.

RONNEBERGER, Olaf, Philipp FISCHER, and Thomas BROX. “U-net: Convolutional networks for biomedical image segmentation”. In: *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer. 2015, pp. 234–241.

RUIZ, Lina and Fabio MARTÍNEZ. “Weakly Supervised Polyp Segmentation from an Attention Receptive Field Mechanism”. In: *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 2022, pp. 3745–3748.

SÁNCHEZ-MONTES, Cristina et al. “Computer-aided prediction of polyp histology on white light colonoscopy using surface pattern analysis”. In: *Endoscopy* 51.03 (2019), pp. 261–265.

SILVA, Juan et al. “Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer”. In: *International Journal of Computer Assisted Radiology and Surgery* 9.2 (2014), pp. 283–293.

SUN, Xinzi et al. “Colorectal polyp segmentation by u-net with dilation convolution”. In: *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. IEEE. 2019, pp. 851–858.

TAJBAKHSH, Nima, Suryakanth R. GURUDU, and Jianming LIANG. “Automated polyp detection in colonoscopy videos using shape and context information”. In: *IEEE Transactions on Medical Imaging* 35.2 (2015), pp. 630–644.

TANAKA, Shinji et al. “Evidence-based clinical practice guidelines for management of colorectal polyps”. In: *Journal of Gastroenterology* 56 (2021), pp. 323–335.

TOMAR, Nikhil Kumar et al. “Fanet: A feedback attention network for improved biomedical image segmentation”. In: *IEEE Transactions on Neural Networks and Learning Systems* (2022).

TOMAR, Nikhil Kumar, Debesh JHA, and Ulas BAGCI. “DilatedSegNet: A Deep Dilated Segmentation Network for Polyp Segmentation”. In: *International Conference on Multimedia Modeling* (2023), pp. 334–344.

VASWANI, Ashish et al. “Attention is all you need”. In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.

WEI, Jun, Shuhui WANG, and Qingming HUANG. “F^3Net: fusion, feedback and focus for salient object detection”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 07. 2020, pp. 12321–12328.

YU, Jieyao et al. “Fully Convolutional DenseNets for Polyp Segmentation in Colonoscopy”. In: *2019 IEEE 35th International Conference on Data Engineering Workshops (ICDEW)*. 2019, pp. 306–311.

*

APPENDIX

appendix A. Academic Products

Journals

- RUIZ, Lina, Luis GUAYACÁN, and Fabio MARTÍNEZ. “Attention Maps to Highlight Potential Polyps during Colonoscopy”. *Tecnura*. 2023.
  Status: Published.
- RUIZ, Lina, Jair RUIZ, and Fabio MARTÍNEZ. “Segmenting polyps in long colonoscopies from a multi-head cross-attention mechanism trained under a weakly supervised framework”. *Medical Image Analysis*. 2023.
  Status: Manuscript prepared for submission.
- RUIZ, Lina, Franklin SIERRA, and Fabio MARTÍNEZ. “Polyp segmentation and classification from a multi-tasks cross-attention strategy”. *SPIE*. 2023.
  Status: Manuscript prepared for submission.

Conference papers

- RUIZ, Lina and Fabio MARTÍNEZ. “Weakly Supervised Polyp Segmentation from an Attention Receptive Field Mechanism”. *44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 2022.
  Status: Presented.