On the Exploitation of Temporal Redundancy to Improve Polyp Detection in Colonoscopy

1st Giovanna Pappalardo
Department of Mathematics and Computer Science
University of Catania
Catania, Italy
giovanna.pappalardo1@unict.it

2nd Dario Allegra
Department of Mathematics and Computer Science
University of Catania
Catania, Italy
allegra@unict.it

3rd Filippo Stanco
Department of Mathematics and Computer Science
University of Catania
Catania, Italy
fstanco@dmi.unict.it

4th Giovanni Maria Farinella
Department of Mathematics and Computer Science
University of Catania
Catania, Italy
Cognitive Robotics and Social Sensing Laboratory
ICAR-CNR
Palermo, Italy
gfarinella@dmi.unict.it

***Abstract*—Colonoscopy is currently the most effective screening method to find precancerous colon polyps and plan their removal. Computer-aided polyp detection can reduce polyp miss detection rates and help doctors find the most critical regions to pay attention to. The challenge in detecting polyps is due to the polyp's morphology and size, and these fall into false-negative. Indeed, polyps may exhibit high variability in shapes (e.g., depressed, flat, pedunculated, etc...). Moreover, the water injected from the endoscope results in artifacts which impede the detection, and the lubricating mucus causes light artifacts due its glossiness. To address this problem, we propose a mask-based attention mechanism to ensure that the employed detector focuses on particular regions of the image in order to reduce misdetection rate. Our contribution takes advantage of information on polyp's position over time within a video sequence. We provide such information through a binary mask which points out the last-known polyp's position. The proposed approach is validated on a dataset that has been labeled by colonoscopy experts. It contains about 200 videos and more than 500 different polyps with high variability in size and textures. Experimental results show that the proposed attention mechanism recover a smaller number of false negatives and achieves an F1-score of 80.21%.**

***Index Terms*—Polyp Detection, Mask, Attention**

I. Introduction

In the last decades computer vision and machine learning algorithms have been massively employed to design and develop systems to support medical tasks. These systems include tools and software for patients monitoring, as well as for clinical and diagnostic purpose. Several studies have been done in this context to support health professionals, such as breast shape analysis for reconstructive surgery [1], [2], CT image analysis for cysts detection [3], semantic segmentation for intraoperative guidance [4], food image analysis for diet monitoring [5], objective evaluation of acne severity [6].
However, multiple medical disorders can appear in the gastrointestinal tract, from simple nuisances to serious diseases which may jeopardise human life. Among them, the Colorectal cancer (CRC) is the second most common cause of cancer-related death for both sexes in many parts of the world. Often, the antecedents of the CRC are polyps that mutate and progress slowly, becoming invasive tumors that metastasizes other parts of the body. Since the risk of cancer development can be reduced by early detection, colonoscopy is employed as primary method for screening and prevention of CRC. However, during a colonoscopy, a significant fragment of polyps can be missed [7], especially the ones located in the proximal colon. Moreover, the screening process is an operator-dependent task; hence, human factors, such as fatigue, insufficient attentiveness during colon examination, and lack of sensitivity to visual characteristics of polyps. Missed polyps cause a survival rate of less than 10% [8]. This motivates the use of computer-aided polyp detection to help colonoscopists to reduce false negatives [9]. In recent years, automatic polyp detection systems using deep-learning methods have been proposed [10], [11], [12] for detecting colorectal polyps in real-time colonoscopy videos. Additional investigations are significant in showing the generalizability of deep-learning algorithms concerning the variations in scale, location, and brightness of polyps.
Even though deep-learning methods generally achieved good performances, they require a high number of training data to successfully work. In this paper, we employ a dataset which includes hundreds of thousands of labeled frames which comes from real colonoscopy videos [13]. Our dataset presents a high variability in polyps texture and morphology, as well as in mucosa appearance. Even with deep-learning techniques trained on a big dataset we notice that many polyps are hard to detect. Indeed, using a standard detection approach, in which we only observe the current frame, we get low Recall and high Precision. The reason is precisely linked to the characteristics

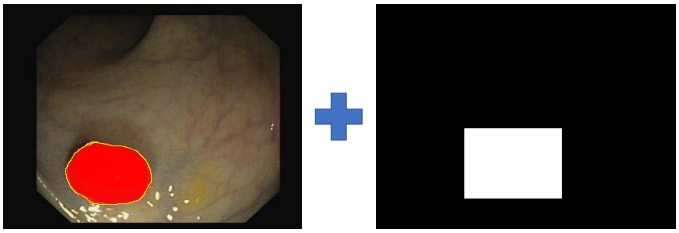

Fig. 1. Attention Mechanism. The procedure for creating input images of each video sequence by the combination of the original frame (left) and Mask related to the previous frame (right).

of the polyps, which cause confusion in detection and fall into false negatives. For example, flat polyps are confused with the mucous membrane or more evident polyps are sometimes confused with probe light or water. Our insight to improve the detector capability is to introduce a sort of attention mechanism which exploits the previous detection to suggest our system to focus in a specific region. This mechanism is based on the realistic assumption which adjacent frames of a videos are similar, hence if a polyp is detected in a certain frame, it could be found in the next one by searching around the same position.

In a nutshell, we exploit temporal properties of video sequences to improve polyps detection. This requires the use of colonoscopies video sequences obtained from real scenarios and labeled by experts. Many works on polyp detection use state-of-art datasets which are small and present low variability. Our contribution consists of an attention mechanism realized with a binary mask which is fed to the detector together with an RGB frame. The binary mask points out the last-known polyp's position in order to give a prior region to easily re-identify a polyp we have already found in the previous frame (Fig. 1).

Experiments, conducted using a modified version of YOLOv3 [14] to take into account the attention mask, prove the validity of the proposed approach which shows better Recall when the attention mechanism is used. It is important to note that we conduct experiments on our dataset only, since state-of-art ones do not include realistic video sequences, which are required to successfully employ the proposed approach.

The remainder of this paper is organized as follows. Section 2 revises the related works. Section 3 shows the details of the proposed method. Section 4 reports the experimental settings, whereas Section 5 discusses the results. Conclusions are summarized in Section 6.

II. Related Work

A. Polyp Detection

Deep learning methods have been applied in this field only in recent years. Previously, classic methods employing swallow features or geometrical properties have been proposed. Hwang et al. [15] proposed a technique in which the polyp region detection is based on the elliptical shape. In [16], texture features are employed for polyps and regular tissue classification. A Support Vector Machine (SVM) is applied as a classification tool in the polyps detection scheme. The authors of [17] employed spatio-temporal features and Conditional Random Field model (CRF). The CRF models the temporal dependences in colonoscopy videos, while multiple eigentissue images, at different angles, robustly model the various tissue types. In addition, the system employs an automatic quality assessment algorithm to preprocess videos by removing low-quality frames. In [18] a method is proposed, which collects a set of edge pixels and then refines this edge map by patch descriptors and classification scheme, before the polyp localization.

The most recent literature is dedicated to the topic of deep learning for the automatic detection of polyps in colonoscopy images. Zhang et al. [10] introduced a novel transfer learning framework utilizing features learned from big nonmedical datasets. This method exploited, in the first step, features of non-polyp images to identify polyp images followed by predicting the polyp histology. Yu et al. [11] designed a novel offline and online three-dimensional deep learning integration framework by leveraging the 3-D fully convolutional network for automated detection of polyps from colonoscopy videos. In [12] it is proposed a system that extracts color wavelet features and convolutional neural network (CNN) features from each sliding window of video frames. The fusion of all the features is fed into SVM for the classification. Dijkstra et al. [19] used a fully convolutional neural network model for semantic segmentation and the transfer learning to produce detection and localization.

B. Visual attention mechanism

There have been some promising works on the visual attention mechanism, but through the development of algorithms applied to different fields from that addressed by us. Xu et al. [20] introduced an attention-based model that automatically learns to describe the content of images; it can show the modality of training in a deterministic manner using standard backpropagation techniques and by stochastically maximizing a variational lower bound. The proposed attention model in [21] not only outperforms average and max-pooling, but it is useful to diagnostically visualize the importance of features at different positions and scales. It introduced extra supervision to the output of fully convolutional neural networks (FCNs) at each scale, and the work proposes to jointly train the attention model and the multi-scale networks. In [22] the authors proposed a novel convolutional neural network called SCA-CNN that incorporates Spatial and channel-wise attention in a CNN. This model learns to pay attention to every feature entry in the multi-layer 3D feature maps. Chu et al. [23] suggested using a visual attention mechanism to automatically learn and infer the contextual representations, driving the model to focus on the region of interest. The approach is proposed

for human pose estimation by stacked hourglass networks to generate attention maps from features at multiple resolutions with various semantics. The conditional random field (CRF) is utilized to model the correlations among neighboring regions in the attention map.

C. Mask attention method

Attention mechanisms have been successfully applied in several contexts. The first part of [24] is related to the introduction of the binary segmentation masks to construct synthetic RGB-Mask pairs as inputs to be used for a mask-guided contrastive attention model (MGCAM) to learn features separately for the person body and background regions. In [25] it is proposed a network composed of two main modules, namely a re-identification (Re-ID) module, and a recurrent mask propagation (Re-MP) module. The Re-ID module helps to build confident starting points in non-successive frames and retrieve missing segments generated by occlusions. Based on the segments provided by the Re-ID module, the Re-MP module propagates their masks bidirectionally from a recurrent neural network to the full video. Authors of [26] exposed both the reference frame with annotation and the current frame with previous mask estimation to a deep network. The network detects the target object by matching the appearance at the reference frame and also tracks the previous mask by referencing the previous target mask in the current frame. Differently by previous works we exploit an attention mechanism for polyp detection based on temporal redundancy. We train the object detector using the current frame and a binary mask which specifies the last known position of the lesion to push the network focus on specific regions of the frame with the aim of reducing false negative rate.

III. Proposed Method

The goal of our method is to perform polyp detection by exploiting temporal redundancy to take advantage of detection masks related to previous frames. Hence, we train an object detector using the current frame together with the mask relating to the position of the polyps in the previous frame when this last information is available.
Therefore, our contribution focuses on modelling an object detector which optimally uses information about previous polyp's position in a video sequence. It allows to build a sort of lesion tracker which exploits the temporal properties of the lesion during the whole screening process. In the next sections we detail the proposed attention mechanism to train a network which may exploit previous detection results through binary masks. Then, we provide a short description of YOLOv3, the CNN architecture used as detector to validate our method.

A. Attention Mechanism by Mask

Let be F_j the RGB $j-th$ frame in a colonoscopy video and let be $M(F)$ a function to assign a binary mask to the ground truth bounding box of the frame F in which 1 indicates a pixel inside a bounding box and 0 a pixel outside it. We propose to train a detector by providing the input pair $(F_j, M(F_{j-1}))$ and

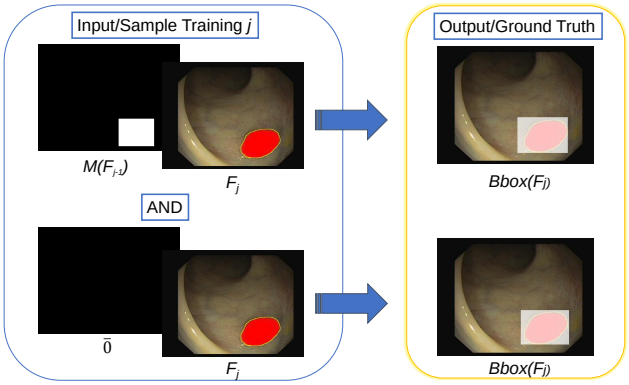

Fig. 2. Construction of the input for the training. Given a frame F_j, it is combined with the binary mask $M(F_{j-1})$. The same Fj is also combined with the mask $\bar{0}$ in order to train the network even when no polyps occurs in the previous frame. The input data of the object detector is a four-channel signal.

the bounding box annotation of the frame F_j (Fig. 2). We train the network by including knowledge on the previous polyps' position. Hence, the input is a $H \times W \times 4$ tensor obtained by merging an RGB image related to F_j and the mask $M(F_{j-1})$. In this paper we employ YOLOv3 [14] as detector, and we change the first layer in order to input a $H \times W \times 4$ tensor in place of a standard RGB image. However, as many frames do not present polyps (negative frames), they drive masks where each element in the mask is 0. For the sake of readability we indicate such mask with the term $\bar{0}$.

To make the network robust, and able to deal even with frame preceded by a negative one, we also train the network with all the pairs $(F_j, \bar{0})$. Hence, frames which present polyps in their previous one, are fed in the network twice, the first time with the proper mask and the second time with $\bar{0}$ mask.

Finally, we assume that the first frame of a sequence is always preceded by a negative frame.

B. The YOLOv3

In this work we choose YOLOv3 architecture as polyps detector, which is the third version of YOLO [27]. It presents better backbone classifier with respect to the first generation and a higher average precision for small objects. The three different scales for the object are obtained by downsampling the size of the input image by 32, 16, and 8, respectively. Also, YOLOv3 uses independent logistic classifiers for each class instead of a regular softmax layer. This architecture has 53 convolutional layers and the first one input layer accepts a 416×416 image. Ground truth annotations for an image are given in text form, by reporting a line for each object which include the centre position (x, y) of the bounding box and its size (i.e, $width$ and $height$). The input image is expected to be an RGB images, namely a $416 \times 416 \times 3$ tensor. However, we change the input layer to make YOLOv3 able to accept $416 \times 416 \times 4$ tensor, in which the new channel includes the binary attention mask.

Fig. 3. Example of our video sequence.

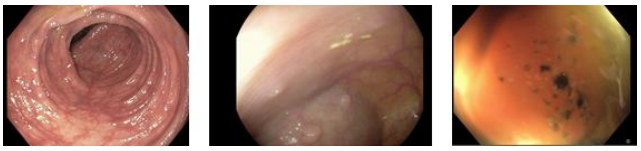

Fig. 4. Variability of the polyps and mucosa in the dataset.

IV. Experimental Settings

In this section, we evaluate polyp detection performance to prove the attention mechanism effectively decreases the number of false negatives. Note that in the considered application context false negative have to be reduced in order to reduce the risk for the patient under consideration in the colonoscopy screening. The experiments are conducted on our dataset made up entirely of real video sequences and labeled by colonoscopy experts. For the performance evaluation, the dataset is split into 70% for the train set and 30% for the test set. We remark our dataset includes over 100 videos and exhibits a high variability in term of scale, illumination, polyp's shape and texture. Some frames do not present any lesion in order to train the model under multiple scenarios.

A. Dataset

In our dataset [13], the same polyp occurs in a video sequence for a large number of consecutive frames (Fig. 3). Of course, sequences which do not present any lesions in a subset of frames are present.
Our contribution is based on exploiting the information of colonoscopy video sequences, which are nothing more than temporal frames.
The dataset has been labeled by experts with ground truth bounding boxes for each polyp. The dataset contains more than 500 different polyps and about 200 videos, and allow us to learn a detector which may exploit temporal information. Among the sources of variability of the polyps in the dataset are the type and occlusions (Fig. 4). The dataset has a high variability in terms of size of polyps. Our idea is not applicable with the datasets available in literature [28], [29], [30] as they often have short sequences with a low frame-rate. On the other hand, our detector is used to train on realistic sequences and can take advantage of the temporal information.

B. Evaluation Metrics

Performances are evaluated by popular metrics: Precision ($Prec$), Recall (Rec) and F1-score ($F1$) [31]. Specifically, correctly identified polyps are considered True Positives (TP). If no polyps are found on an image without a lesion, the result is considered True Negative (TN). A False Positive (FP) occurs when a polyp is incorrectly detected on a normal mucosa. Finally, a False Negative detection (FN) occurs when a polyps which appears in a frame is not detected.

C. Experiments

The experiments are carried out taking into account different cases. We test our approach on 30% of the dataset, whereas 70% of frames are used for training. The split is performed such that all frames of a video belong to only one of the two (i.e., training and test set do not contain frames of the same video). We consider the following tests:

- Baseline test: to evaluate the performance of the model D trained with no attention mask. Hence, the model is trained by input the RGB frame only, as in a standard detector;
- Temporal test: to evaluate the performance achieved by training the model DM which uses the attention mask. Detection of the DM model are combined with the model D for the final detection.

The baseline test is useful for comparing a standard detection approach with respect to the proposed framework. The second test highlights the benefits of the proposed approach, in which detection result at frame $j-1$ is given as input fot the detection at frame j. For the final detection, we combine the two different detectors, namely the the one trained with RGB only (D) and the one trained with RGB and attention mask (DM). Fig. 5 shows a flowchart which describes how the model D and DM are combined: the input frame F_j is fed in the system for the inference, then we check the mask related to the previous detection $\tilde{M}(F_{j-1})$. If no polyps are detected in the previous frame, no attention mask is provided ($\tilde{M}(F_{j-1})$ = $\bar{0}$) and the RGB frame F_j is fed in the standard detector D. Otherwise, if the system detects a polyps in the previous frame F_{j-1}, we concatenate the binary mask $\tilde{M}(F_{j-1})$ and the RGB frame F_j and input them to the model DM. The process is repeated over time along the overall video sequences. This test is performed by using as attention mask the detection result of previous steps. This means if the system results in a false positive or a false negative, a tricky attention mask could be input in the next step.

To check the performance with an oracle which know always the polyp's position at the previous step, we perform a third test in which the inferred masks $\tilde{M}$ are replaced with the ground truth mask M. This test gives us the best performance which our strategy could achieve.

V. Results

In this section we discuss experimental results achieved with the proposed framework. In Table I we report a quantitative evaluation in term of Precision ($Prec$), Recall (Rec) and F1-score ($F1$). The comparison between the standard detector (D) and the proposed one (DM + D), exhibits a Recall

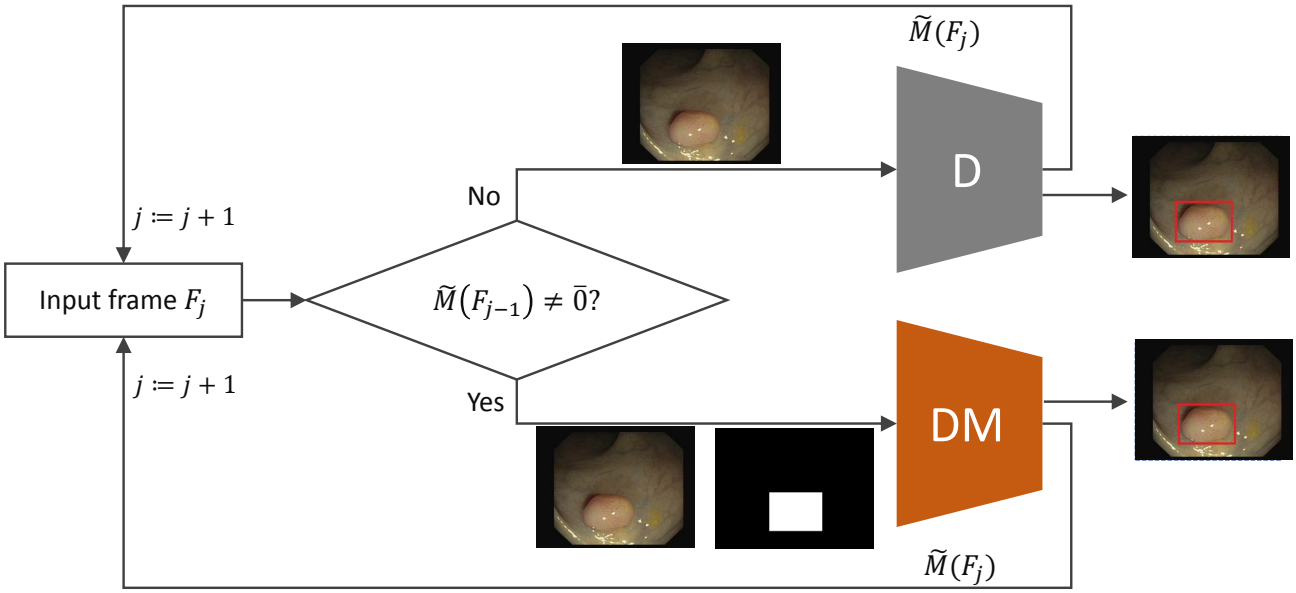

Fig. 5. Combination of the two detectors. The sample F_j is given to the system and the attention mask $\tilde{M}(F_{j-1})$ is checked. If it is available ($\tilde{M}(F_{j-1}) = \bar{0}$), F_j and the related mask are concatenated and input into the model DM. Else, if no attention mask is available, only the frame F_j is fed to the standard model D.

TABLE I
EVALUATION COMPARISON

Detector	Prec	Rec	F_1
D	95.02	64.53	76.86
D + DM	95.54	69.16	80.21
D + DM with GT	93.60	87.67	90.53

improvement of 4.63%, which means we decrease false negative rate. In few words, the proposed attention mechanism, which exploits temporal redundancy, reduces the misdetection of polyps.

In addition, the proposed approach achieves a sligth improvement even on Precision ($+0.52\%$), which means it decreases the false positive rate (normal mucosa incorrectly identified as lesion). Of course, this led a raise of F1-score ($+3.35\%$) as it combines Recall and Precision.

Finally, we report results obtained in the ideal scenario in which the attention mask is always correctly built. In this case the system always know the polyp's position in the previous frame and uses it the perform the detection at the current frame. With this oracle the Recall improvement is reasonably higher ($+23.14\%$). Despite experiments were conducted with YOLOv3, it is possible to use any object detector together with the proposed attention mechanism.

Different error analysis are carried out to verify the usefulness of our contribution and to understand which frames can be recovered thanks the proposed strategy. The analysis focuses on false negatives obtained from the experiments carried out on our dataset with the base detector YOLOv3. Initially, video sequences with the highest percentage of error are analyzed, and then we pay attention on the false negatives. Specifically, we focus on the misdetection of the model D and we found that about 11% of false negative presented a true positive in the previous frame. This means our method improves the Recall by mainly operating on such false negative frames, which are recovered by exploiting the attention mask which comes from the polyps of the previous frame correctly detected.

VI. CONCLUSION

In this work, we propose a simple attention mechanism to be integrated with an object detector to improve the performance of polyp detection. The main idea is to exploit temporal redundancy and improve the detection by using previously detected polyps. Experimental results, conducted by using YOLOv3 detector, confirm that the proposed approach we obtain an improvement of 4.63% in Recall, by decreasing the misdetection. This approach can be used in a real context in real-time colonoscopies since the temporal redundancy of the data is exploited.

Future works can be devoted on extending the current approach by learning attention mask explicitly for the considered task.

ACKNOWLEDGMENT

This research is partially supported by Piano della Ricerca 2016-2018 - CHANCE - Linea di Intervento 1 of DMI, University of Catania.

REFERENCES

[1] G. M. Farinella, G. Impoco, G. Gallo, S. Spoto, G. Catanuto, and M. B. Nava, "Objective outcome evaluation of breast surgery," in *Proceedings of the 9th International Conference on Medical Image Computing and Computer-Assisted Intervention - Volume Part I*, ser. MICCAI'06, 2006, pp. 776–783.

[2] G. Gallo, D. Allegra, Y. G. Atani, F. L. M. Milotta, F. Stanco, and G. Catanuto, "Breast shape parametrization through planar projections," in *Advanced Concepts for Intelligent Vision Systems*, 2016, pp. 135–146.

[3] S. Battiato, G. M. Farinella, G. Gallo, O. Garretto, and C. Privitera, "Objective analysis of simple kidney cysts from ct images," in *2009 IEEE International Workshop on Medical Measurements and Applications*, 2009, pp. 146–149.

[4] D. Ravì, H. Fabelo, G. M. Callic, and G. Yang, "Manifold embedding and semantic segmentation for intraoperative guidance with hyperspectral brain imaging," *IEEE Transactions on Medical Imaging*, vol. 36, no. 9, pp. 1845–1857, 2017.

[5] D. Allegra, M. Anthimopoulos, J. Dehais, Y. Lu, F. Stanco, G. M. Farinella, and S. Mougiakakou, "A multimedia database for automatic meal assessment systems," in *Lecture Notes in Computer Science*, vol. 10590, 2017.

[6] A. Melina, N. N. Dinh, B. Tafuri, G. Schipani, S. Nisticò, C. Cosentino, F. Amato, D. Thiboutot, and A. Cherubini, "Artificial intelligence for the objective evaluation of acne investigator global assessment," *Journal of Drugs inDermatology*, vol. 17, no. 9, pp. 1006–1009, 2018.

[7] J. C. Van Rijn, J. B. Reitsma, J. Stoker, P. M. Bossuyt, S. J. Van Deventer, and E. Dekker, "Polyp miss rate determined by tandem colonoscopy: a systematic review," *American Journal of Gastroenterology*, vol. 101, no. 2, pp. 343–350, 2006.

[8] L. Rabeneck, H. B. El-Serag, J. A. Davila, and R. S. Sandler, "Outcomes of colorectal cancer in the united states: no change in survival (1986–1997)," *The American journal of gastroenterology*, vol. 98, no. 2, pp. 471–477, 2003.

[9] A. Pabby, R. E. Schoen, J. L. Weissfeld, R. Burt, J. W. Kikendall, P. Lance, M. Shike, E. Lanza, and A. Schatzkin, "Analysis of colorectal cancer occurrence during surveillance colonoscopy in the dietary polyp prevention trial," *Gastrointestinal endoscopy*, vol. 61, no. 3, pp. 385–391, 2005.

[10] R. Zhang, Y. Zheng, T. W. C. Mak, R. Yu, S. H. Wong, J. Y. Lau, and C. C. Poon, "Automatic detection and classification of colorectal polyps by transferring low-level cnn features from nonmedical domain," *IEEE journal of biomedical and health informatics*, vol. 21, no. 1, pp. 41–47, 2016.

[11] L. Yu, H. Chen, Q. Dou, J. Qin, and P. A. Heng, "Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos," *IEEE journal of biomedical and health informatics*, vol. 21, no. 1, pp. 65–75, 2016.

[12] M. Billah, S. Waheed, and M. M. Rahman, "An automatic gastrointestinal polyp detection system in video endoscopy using fusion of color wavelet and convolutional neural network features," *International journal of biomedical imaging*, vol. 2017, 2017.

[13] G. Pappalardo and G. M. Farinella, "On the detection of colorectal polyps with hierarchical fine-tuning," in *2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. IEEE, 2020, pp. 1–5.

[14] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[15] S. Hwang, J. Oh, W. Tavanapong, J. Wong, and P. C. De Groen, "Polyp detection in colonoscopy video using elliptical shape feature," in *2007 IEEE International Conference on Image Processing*, vol. 2. IEEE, 2007, pp. II–465.

[16] D.-C. Cheng, W.-C. Ting, Y.-F. Chen, Q. Pu, and X. Jiang, "Colorectal polyps detection using texture features and support vector machine," in *International Conference on Mass Data Analysis of Images and Signals in Medicine, Biotechnology, and Chemistry*. Springer, 2008, pp. 62–72.

[17] S. Y. Park, D. Sargent, I. Spofford, K. G. Vosburgh, A. Yousif *et al.*, "A colon video analysis framework for polyp detection," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, pp. 1408–1418, 2012.

[18] N. Tajbakhsh, C. Chi, S. R. Gurudu, and J. Liang, "Automatic polyp detection from learned boundaries," in *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2014, pp. 97–100.

[19] W. Dijkstra, A. Sobiecki, J. Bernal, and A. Telea, "Towards a single solution for polyp detection, localization and segmentation in colonoscopy images," in *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2019, pp. 616–625.

[20] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[21] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3640–3649.

[22] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "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.

[23] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1831–1840.

[24] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1179–1188.

[25] X. Li and C. Change Loy, "Video object segmentation with joint re-identification and attention-aware mask propagation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 90–105.

[26] S. Wug Oh, J.-Y. Lee, K. Sunkavalli, and S. Joo Kim, "Fast video object segmentation by reference-guided mask propagation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7376–7385.

[27] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[28] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015.

[29] J. J. Bernal, A. Histace, M. Masana, Q. Angermann, C. Sánchez-Montes, C. Rodriguez, M. Hammami, A. Garcia-Rodriguez, H. Córdova, O. Romain, G. Fernández-Esparrach, X. Dray, and J. Sanchez, "Polyp Detection Benchmark in Colonoscopy Videos using GTCreator: A Novel Fully Configurable Tool for Easy and Fast Annotation of Image Databases," in *Proceedings of 32nd CARS conference*, 2018.

[30] J. Bernal and H. Aymeric, "MICCAI Endoscopic Vision Challenge Polyp Detection and Segmentation," https://endovissub2017-giana.grand-challenge.org/home/, online; accessed 01 May 2019.

[31] A. I. Bandos, H. E. Rockette, T. Song, and D. Gur, "Area under the free-response roc curve (froc) and a related summary index," *Biometrics*, vol. 65, no. 1, pp. 247–256, 2009.