UNIVERSITÀ DEGLI STUDI DI CATANIA

Dipartimento di Matematica e Informatica

Dottorato di Ricerca In Informatica XXXIV Ciclo

Giovanna Pappalardo

Polyp Detection in Colonoscopy Video

Tesi di Dottorato di Ricerca

Tutor: Prof. Giovanni Maria Farinella

Anno Accademico 2020 - 2021

Declaration of Authorship

I, Giovanna Pappalardo, declare that this thesis titled, "Polyp Detection in Colonoscopy Video" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Abstract

Colonoscopy is the only screening that can detect polyps in the colon that can mutate into colorectal cancer. The task of detecting polyps is a job done by the doctors themselves who perform the procedure. It often happens that polyps are not recognized for reasons such as fatigue of the doctor himself, preparation in the previous days for the procedure not done correctly by the patient, confusion between polyp or mucosa, probe lights, or water injected. The work in this thesis was carried out thanks to a grant from Linkverse s.r.l., a company based in Rome, Italy, which believes that Machine Learning can help physicians in the detection of polyps and thus can help them identify the most difficult ones. During the Ph.D. course, we tried to solve the problems related to this topic by using known Object Detection architectures. The first contribution is related to the specialization of a known object detection network through clustered features and fine-tuning steps. The second contribution concerns an attention mechanism integrated into an object detection network to focus on specific regions of each image used in the training phase. Finally, the last contribution is about a framework created for Data Augmentation by exploiting a known Inpainting network. This contribution is useful to provide other researchers with a more extensive and variable dataset with realistic data.

Acknowledgements

I would like to express my gratitude to my PhD supervisor, Prof. Giovanni Maria Farinella for being a guide during these three years, in which his support has been fundamental. Likewise, I would also like to thank Dr. Dario Allegra for his advice. I would also like to thank all the professors and colleagues in the IPLAB group for their suggestions during my PHD studies.

I would like to offer my special thanks to my partner Giuseppe and my family for their constant support.

Contents

Chapter 1

Introduction

1.1 Motivation

Among the medical tasks, colonoscopy aims at the screening of the intestine to detect lesions (e.g., polyps) and to plan their removal. Nowadays, it is the gold standard for diagnosis of colon cancer (CRC). Often, the antecedents of the CRC are polyps that mutate and progress slowly, becoming invasive tumors that metastasizes other parts of the body. A colon polyp, which can be seen in the figure 1.1, is a cluster of cells found within the colon when developed within the colon and often appear as a small hill like structure [1]. Since the risk of cancer development can be reduced by early detection, colonoscopy is employed as primary method for screening and prevention of CRC [2]. However, the colonoscopy procedure suffers of an high percentage of miss-detections of polyps during the screening, especially in the early stage of the disease. More recently, computer-aided diagnosis (CAD) and automated Computer Diagnosis (ACD) were born in this context and contribute in helping physicians make the process more automated. This is possible due in part to the success of deep learning for the polyp detection task. 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. 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. The doctor may also perform a polypectomy (removal of a polyp) during the procedure if the polyp is under a certain size, otherwise surgery is used later [3]. Screening has also assumed a key role in Italy between the ages of 50 and 69 years. In the years between 2003 and 2014, incidence and mortality rates decreased significantly in all areas except in the south and islands, where incidence increased and mortality remained stable [4]. Missed polyps cause a survival rate of less than 10% [5]. The aforementioned considerations motivate the introduction of new technologies able to improve the rate of identification of intestinal lesions.

In addition, the state of the art analysis allowed us to infer that the datasets available to researchers in this area are limited in terms of variability. The absence of large and diverse dataset in this context does not allow an adeguate comparison of the algorithms for polyp detection. Many studies use images in which the polyps appear mainly/only in the foreground and in most of the cases in the center of the images [6]. Moreover, images are usually extracted by medical devices, so depth of color is compromised and the compression can destroy useful information to be exploited by a detector. Another problem is that many of the images composing the current

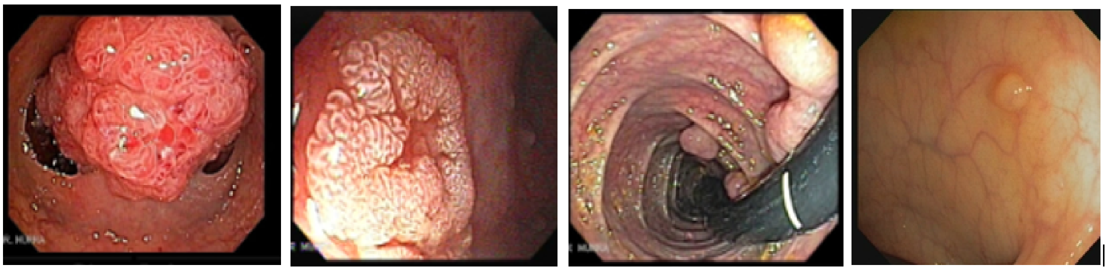

Figure 1.1: Different polyps in coloscopies frames.

available datasets are related the same polyp (i.e., samples of consecutive frames of a colonoscopy video related to a polyp which have very similar appearance). Ideally, a benchmark dataset should contains thousand of images related to a very high number of different polyps. Therefore, in addition to the aspect of Computer-aided polyp detection it was also found important to generate Colonoscopy Data in order to increase the data but at the same time addressing new methodologies applied to this field. Still in this field the perfermonces are not fully acceptable unlike other types of detection such as in other application domains. This is mostly due to the lack of labeled colonoscopy datasets available and labeled by experienced clinicians. Other issues with datasets relate to variability and label type. Often it happens in fact to have to train small datasets or medium / large size but with little variability in terms of shape, color or scale. The labels that are provided are often segmentation masks that do not respect the shape of the polyp itself but rather localize the lesion in the image through ellipses or round shapes, so there is no real label of the data. The need to have large datasets allows then to generalize a training and have better performance in detection.

1.2 Objectives and Approaches

This thesis work focuses on the detection of polyps in colonoscopies. Detection in the medical field is indeed a difficult and critical, but it helps to speed up the time of detection of lesions that may mutate into colorectal cancer. So detection in this context is useful to the endoscopist. The images used are in fact derived from real video sequences that are analyzed by a learning algorithm to identify a polyp inside a normal mucosa. Therefore, this work does not replace the work of a doctor, but it is a support to refine the level of precision that can escape the human eye. This study has allowed us to identify important aspects in this field of research:

- the lack of real and numerous data;
- the difficulty in detecting some types of polyps due to artifacts and thus in having many false negatives with the use of neural networks;
- the similarity between normal mucosa and polyp.

The goal of the research initiated in the first year of the Ph.D. program was to improve the detection of intestinal lesions in terms of timing and recognition of polyps that are the cause of intestinal tumors and overcome the limits described above. Initially, the company dataset was carefully studied to clean up the data and analyze issues in the frames themselves that could interfere with any object detector used. Therefore, a study of the state of the art was carried out in order to search for existing methods for the detection of polyps in colonoscopies and gastroscopies, but also of Object Detection methods based on Deep Learning. Considering also the study of the state of the art, polyp-detection through YoloV3 [7] has been carried out and results have been compared with the state of the art and with company results obtained in the past; in addition, an error analysis has been conducted and a specialization of an architecture based on problems detected by the error analysis has been created. Specifically, in order to improve the detection performance it was decided to proceed with the study of morphologies and textures related to the mucosa in the frames and to create clusters. The neural network was specialized according to the morphological characteristics of the mucosa of each subset by performing a Fine-Tuning phase [8].

Errors were then analyzed and it was found that many false negatives also fall within the same video sequence between true-positives, so it was decided to exploit this analysis to create a method to exploit information between frames within a video sequence to detect as many true-positives as possible. It was also decided to create an attention mechanism for the detector by exploiting the mask of the frame preceding the considered frame.

Another topic of particular interest is data generation in this domain. It was decided to use not corporate data, but available open-source data. Already from the experiments with detectors, it has been found that using these datasets is more difficult in the detection of polyps precisely because of the low numerosity of the data and unrealistic characteristics for the purpose of an accurate research study. At the state of the art, there are works that mainly exploit Generative Adversarial Networks (GANs) [9] or Conditional GANs [10] to generate realistic data from small publicly available datasets. Therefore, a number of scientific publications on the topic have been explored in depth. In the literature for data generations is now used also the Inpainting technique, so also related works on this have been deepened. A data augmentation framework based on inpainting was then proposed to generate realistic data.

1.3 Contributions

The main contributions of this thesis are the follow:

- analysis of existing methodologies of object detection for polyp detection;
- analysis of existing open-source datasets and industrial dataset to learn the issues;
- feature extraction for polyps to create clusters or specialization;
- deep learning method specialization based on features of the polyp;
- analysis of results with detectors to detect issues within video sequences;
- creation of an attention mechanism to improve polyp detection method;
- investigation of open-source datasets to understand the missing features and then how to generate new data to improve the training performance in this field;
- study of common technique of inpainting;
- proposal of a framework for data augmentation using inpainting.

The principal contribution of this thesis submitted and published in international conferences:
International Journal:

- Pappalardo, G., Allegra, D., Stanco, F., & Farinella, G. M. Inpainting Based Data Augmentation to Improve Polyp Detection in Colonoscopy. Submitted to Computers in Biology and Medicine (CBM) Journal.

International conferences:

- Pappalardo, G., & Farinella, G. M. (2020, June). On the Detection of Colorectal Polyps with Hierarchical Fine-Tuning. In 2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA) (pp. 1-5). IEEE.
- Pappalardo, G., Allegra, D., Stanco, F., & Farinella, G. M. (2020, December). On the Exploitation of Temporal Redundancy to Improve Polyp Detection in Colonoscopy. In 2020 IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS) (pp. 58-63). IEEE.

1.4 Thesis Outline

The remainder of this Ph.D. Thesis is divided in 6 chapters related to improve performance for polyp detection. Chapter 2 is a presentation of literature works which have been used for the work in this thesis. Chapter 3 presents neural networks and methods used in this study. Chapter 4 describes a first approach to specialize a neural network by hierarchy. Chapter 5 shows the visual attention mechanism used to improve polyp detection performances. Chapter 6 presents a method of augmenting the data to have a larger and more variable dataset on which to train object detectors for polyp detection. Finally Chapter 7 concludes the thesis and gives insights for future directions. Appendix A reports on work not directly related to this thesis that was published during my Ph.D.

Chapter 2

Related Work

This chapter reports related work from the major themes addressed during the study. Specifically reported are the state-of-the-art contributions that were useful in furthering the topics and providing our contributions.

2.1 Polyp Detection

The recent research focuses on the design and development of automatic polyps detection in colonoscopy videos. Several approaches have been proposed to address this task. A survey on different methods is given in [11]. In [12] a combination of boosting and active learning is exploited to detect regions of interest containing polyps. In [13] multiple hand-crafted features based on color, discrete cosine transformation (DCT) [14] and LBP have been used to feed a fuzzy or a decision tree classifier. In [15] a convolutional neural network called ResYolo with temporal information incorporated via a tracker is proposed to detect polyps and perform temporal refinement of the results. Deep learning vs hand-crafted based methods have been compared in [16] on different publicly available datasets.

In this context, it is very important to identify the largest number of polyps as soon as possible (i.e., ideally during the first screening and, in any case, at very small size). The main limit of the current literature is due to the scarce availability of datasets with the adequate size and variability required to perform a valid quantitative and qualitative evaluation benchmark of different polyp detection algorithms. The absence of large and diverse dataset in this context does not allow an adequate comparison of the algorithms for polyp detection. Many studies use images in which the polyps appear mainly/only in the foreground and in most of the cases in the center of the images [6]. Moreover, images are usually extracted by medical devices, so depth of color is compromised and the compression can destroy useful information to be exploited by a detector.

Another problem is that many of the images composing the current available datasets are related the same polyp (i.e., samples of consecutive frames of a colonoscopy video related to a polyp which have very similar appearance). Ideally, a benchmark dataset should contains thousand of images related to a very high number of different polyps. A large dataset is useful also to properly train polyps detectors based on deep learning architectures. Indeed, considering the available datasets, to deal with training problems when exploiting deep learning based frameworks for polyps detection, some works propose to augment data during training by generating synthetic samples with generative adversarial network (GAN) approaches [16].
Previously, classic methods employing swallow features or geometrical properties have been proposed. Hwang et al.[17] proposed a technique in which the polyp region detection is based on the elliptical shape. In [18], texture features are employed for polyps and regular tissue classification. A Support Vector Machine (SVM) [19] is applied as a classification tool in the polyps detection scheme. The authors of [20]

employed spatio-temporal features and Conditional Random Field model (CRF). The CRF models the temporal dependencies 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 [21] 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.[22] 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.[23] 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 [24] 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.[25] used a fully convolutional neural network model for semantic segmentation and the transfer learning to produce detection and localization.

In [26] the authors used convolutional neural networks (CNNs) combined with autoencoders. To validate the results, they use the three publicly available databases, namely CVC-ColonDB, CVC-ClinicDB and ETIS-LaribPolypDB [27, 28, 6]. They applied classical data augmentation techniques. Mori et al. [29] review and focus on published prospective studies in which AI tools have been used in real-time during colonoscopies to illustrate the practical potential and drawbacks of this cutting-edge technology. They have also explored the potential for clinical implementation of this technology by assessing the regulatory approval requirements for new AI tools for colonoscopies.

2.2 Data Augmentation

Common open source strategies are used to perform data augmentation on datasets. There are two different macro-areas of data augmentation [30]: data augmentations based on basic image manipulations (geometric and photometric transformations), and data augmentations based on Deep Learning techniques. The first group includes geometric transformations (such as rotation, traslation, etc) and color space transformations (such as color jittering). Photometric transformations includes kernel filters, mixing images, random erasing. The last macro-area we investigated in this paper, includes feature space augmentation, adversarial training, GAN-based augmentation, neural style transfer, and meta-learning schemes. Deep based augmentation has been exploited to generate images of colonoscopies with polyps. An example of the classic data augmentation is employed in [31] where the authors perform Polyp Segmentation through the Mask R-CNN. Augmentation is mainly proposed to prevent the problem of overfitting. the The author suggest vertical flipping, horizontal flipping, random rotation between -45 and 45 degrees, arbitrary scaling ranging from 0.5 to 1.5, random shearing between -16 and 16 degrees, random Gaussian blurring with a sigma of 3.0, random contrast normalization by a factor of 0.5 to 1.5, random brightness ranging from 0.8 to 1.5, and random cropping and padding by 0-25% of height and width. Nguyen et al. [32] suggested another methodology to improve the segmentation performance. They propose a novel encoder-decoder architecture which can recover the full-resolution prediction by applying data-dependent upsampling method, namely MED-Net, to extract the most useful visual features from multi-scale image inputs. They introduce a new boundary-focused data augmentation method for randomly generating a high number of boundary-aware polyp patterns from each training image. This method contributes to the improvement of MED-Net. They also propose a new adaptive weighted loss function to boost the segmentation performance of MED-Net. They present an attention-based loss function that allows the network to focus more on the polyps and their boundaries. The combination of the loss functions leads to a better performance, because the network can focus on learning iteratively polyp boundaries. Mathew et al. [33] proposed, to expoloit a well known deep learning framework, based on CycleGAN [34]. It is an extended and directional CycleGAN for lossy image-to-image translation unpaired between optical colonoscopy (OC) and virtual colonoscopy (VC). Translating between OC and VC can be generalized to image-to-image domain translation. In their work, authors augment OC video sequences with scale-consistent depth information from VC and augment VC with patient texture, color, and specular reflections from OC (e.g., for realistic polyps synthesis). They introduce a novel extended cycle consistency loss for lossy image domain translation. The network so does not need to hide information in the lossy domain by replacing OC comparisons with VC comparisons. Stronger deletion of these specular reflections and textures are handled via a Directional Discriminator that differentiates the direction of translation as opposed to the standard CycleGAN which is direction-agnostic. This Directional Discriminator is similar to a discriminator in a conditional GAN and deals with paired data thus giving the network, as a whole, a better understanding of the relation between the two domains.

Authors of [35] exploited of generative adversarial networks (GANs) for synthetic data generation. For classification, a CNN is trained to discriminate between benign polyps vs. malignant polyps. Authors conducted experiments on two datasets demonstrating that the GAN based data augmentation technique can be effectively used to improve colonic polyps classification.

2.3 Inpainting

Differently, in the present work we propose to improve detection performance by operate a data augmentation through an inpainting approach. Arnold et al. [36] a method for segmentation of specular highlights based on nonlinear filtering and colour image thresholding and an efficient inpainting method that alters the specular regions in a way that eliminates the negative effect on most algorithms and also gives a visually pleasing result. They also present an application of these methods in improvement of colour channel misalignment artefacts removal. Their inpainting algorithm is performed on two levels. They first use the filling technique where they modify the image by replacing all detected specular highlights by the centroid colour of the pixels within a certain distance range of the outline, then they filter this modified image using a Gaussian kernel. For the second level, the binary mask marking the specular regions in the image is converted to a smooth weighting mask. In [37] the authors present a work focused on the development of automatic polyp localization methods. They present the first study that takes into account the impact of different endoluminal scene elements in polyp localization results. They address the influence of specular highlights, blood vessels and the black mask that surrounds the endoluminal scene. Their method integrates valley information to locate the polyp.

They discern between valley information that comes from polyps and the one that is related to other elements in order to improve polyp localization results. The novelty of their work presented is the assessment of the impact that different elements of the endoluminal scene have on polyp localization results. In this regard, they exploited inpainting and, precisely, a diffusion step of the inpainting algorithm for a pixel: they considered pixels under a given detection mask M and pixels outside M. Then, a calculation of the new value from the valid neighbors is performed. To get an even more realistic image, they created a dilated mask by performing a dilation with a circular structural element and then convolving the result with a Gaussian kernel.

Alsaleh et al. in [38] proposed a mirror region segmentation method based on an automatic color matching threshold and a gradient-based edge detector. Their insight is that specular reflections are common in endoscopic images and such reflections are caused by the strong reflectivity of the mucus layer on the organs and the relatively high intensity of the light source. This problem is a source of error that can affect the performance of screening and any other system for polyp detection. Segmented regions are recovered using a robust mask-specific Sobolev inpainting approach [39], corresponding to interpolating missing pixels using surrounding information.

Akbari et al. [40] address the problem of detection and removal of reflections, in order to improve the image quality of colonoscopy and facilitate the diagnosis procedure. They propose a novel reflection detection method based on both RGB and HSV color spaces with an SVM classifier. They also introduce an inpainting method based on patch selection around each reflection region. It consists of appropriate selection of replacement patches and removal of blocking effects They also propose an edge smoothing algorithm to enhance the quality of inpainted image.

2.4 Dataset

One of the main problem in assessing the performances of polyp detection framework is the lack of publicly available and representative dataset to perform experiments. The most popular dataset available is the CVC-12k [41, 42]. It is composed by only 11,954 images with ground truth masks related to only 18 different videos (i.e., 18 patients). Most importantly, images of this dataset are very redundant (more 10K images from only 18 videos) and have small resolution (384 × 288 pixels), which means that some important features that could be useful during training to set parameters of a deep learning based detector to distinguish polyps tissues from normal mucosa could have been destroyed in the resizing. Other dataset in literature in this context, such as CVC-356 and CVC-612 [27] are much smaller than CVC-12k and have lesions which are clearly visible and mostly of them are centered in the frames. In the CVC-612 all images were extracted from 31 different colonoscopy videos which contain 31 unique polyps. All ground truths of polyp regions were annotated by skilled video endoscopists. In all the previous mentioned datasets there is low variability in terms of polyps type (i.e., flat vs others) and few polyps between folds.

Due to the above limitations, for the experiments of this paper we have created a novel dataset twenty times larger than CVC-12k (i.e., $> 200k$ images), with images sampled and labelled by colonoscopy experts considering more than 180 videos. In our company's dataset [43], the same polyp occurs in a video sequence for a large number of consecutive frames. Of course, sequences which do not present any lesions in a subset of frames also included.

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. The dataset has a high variability in terms of size of polyps.

Chapter 3

Preparative Background

In this chapter we provide an overview of the key concepts and known state-of-the-art methodologies used and explored over the course of the PhD years.

3.1 Data Sampling

Data sampling provides a collection of techniques that transform a training dataset in order to balance or better balance the class distribution. Once balanced, standard machine learning algorithms can be trained directly on the transformed dataset without any modification. This allows the challenge of imbalanced classification, even with severely imbalanced class distributions, to be addressed with a data preparation method (Fig. 3.1). Two approaches to make a balanced dataset out of an imbalanced one are under-sampling and over-sampling. Undersampling methods delete or select a subset of examples from the majority class. Under-sampling balances the dataset by reducing the size of the abundant class. This method is used when quantity of data is sufficient. On the contrary, oversampling instead is used when the quantity of data is insufficient. It tries to balance dataset by increasing the size of rare samples. These methods duplicate examples in the minority class

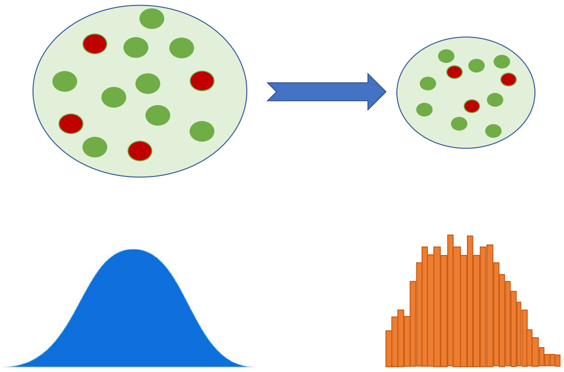

Figure 3.1: Representation of data sampling, which provides a set of techniques that transform a training dataset in order to better balance or equilibrate the class distribution.

or synthesize new examples from the examples in the minority class. Some of the more widely used and implemented oversampling methods include [44]: Random Oversampling, Synthetic Minority Oversampling Technique (SMOTE) [45]. Some of the more widely used and implemented undersampling methods include: Random Undersampling [46], Condensed [47], edited nearest neighbors (ENN) rule [48], Near Miss Undersampling (NMU) [49], One-Sided Selection (OSS) [50], Neighborhood Cleaning Rule (NCR) [51]. A combination of over- and under-sampling is often successful as well. Although an oversampling or undersampling method when used alone on a training dataset can be effective, experiments have shown that applying both types of techniques together can often result in better overall performance of a model fit on the resulting transformed dataset. Some of the more widely used and implemented combinations of data sampling methods include: SMOTE and Random Undersampling, SMOTE and Tomek Links, SMOTE and Edited Nearest Neighbors Rule.

3.2 Clustering: K-Means

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the goal is to separate groups with similar traits and assign them into clusters. There are different types of clustering: exclusive clustering, overlapping clustering, hierarchical clustering. Each methodology follows a different set of rules to define the " similarity" between data points. In fact, more than 100 clustering algorithms are known. But few algorithms are commonly used, let's examine them in detail:

- Connectivity models: as the name suggests, these models are based on the notion that the closest data points in the data space show more similarity to each other than the most distant data points. These models can follow two approaches. In the first approach, they begin by classifying all data points into separate clusters and then aggregating them as the distance decreases. In the second approach, all data points are classified as a single cluster and then partitioned as the distance increases. In addition, the choice of distance function is subjective. These models are very easy to interpret but lack scalability for handling large data sets. Examples of these models are the hierarchical clustering algorithm and its variants.

- Centroid models : these are iterative clustering algorithms in which the notion of similarity is derived from the proximity of a data point to the centroid of the clusters. The K-Means clustering algorithm is a popular algorithm that falls into this category. In these models, the number of clusters required at the end must be mentioned in advance, which makes it important to have prior knowledge of the dataset. These models are run iteratively to find local optimums.

- Distribution models: these clustering models are based on the notion of how likely it is that all data points in the cluster belong to the same distribution (e.g.: normal, Gaussian). These models often suffer from overfitting. A popular example of these models is the expectation maximization algorithm that uses multivariate normal distributions.
- Density models: these models search for areas of different density of data points in the data space. It isolates various regions of different densities and assigns data points within these regions in the same cluster. Popular examples of density models are DBSCAN and OPTICS.

K-means clustering is a type of unsupervised learning, which is used when we have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K (Fig. 3.2). The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. There are two possible ways to choose the optimal number of K clusters. 1) Elbow method: a curve is drawn between WSS (within the sum of squares) and the number of clusters. (2) Purpose-based: run the k-means clustering algorithm to get different clusters based on a variety of purposes. It is possible to see how well they perform for that particular case. The K-means algorithm works in the following way:

- Initialization: randomly initialize two points called cluster centroids. The k-means clustering algorithm is an iterative algorithm and iteratively follows the next two steps.
- Cluster assignment: in this step, perform the calculation of the distance between the initial centroid points with other data points. It means to group the data points which are closer to centriods.
- Move the centroid: calculate the mean values of the clusters created and the new centriod values will these mean values and centroid is moved along the graph.
- Optimization: again the values of euclidean distance is calculated from the new centriods.
- Convergence: finally, this process has to be repeated until find a constant value for centroids and the latest cluster will be considered as the final cluster solution.

Summarizing:

0. Start with initial guesses for cluster centers (centroids)
1. For each data point, find closest cluster center (partitioning step). Write $x_i = (x_{i1}, ...x_{ip})$: If centroids are $m_1, m_2, ...m_k$, and partitions are $c_1, c_2, ...c_k$, then one can show that K-means converges to a *local* minimum of

   $$\sum_{k=1}^{K} \sum_{i \in c_k} ||x_i - m_k||^2 \qquad \text{Euclidean distance}$$

   (within cluster sum of squares)

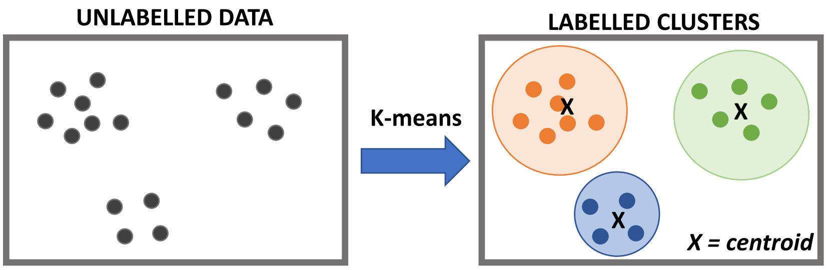

Figure 3.2: Representation of the k-means objective: determining internal clustering in an unlabeled data set.

2. Replace each centroid by average of data points in its partition;

3. Iterate steps 1) and 2) until convergence. continue repeating steps 2 and 3 until the centroids do not change, i.e. a point of convergence is reached where there are no more changes in the clusters.We say that the stop condition in this case has been reached. Usually it is represented by one of the following options: no data points change clusters; the sum of distances is reduced to a minimum; a maximum number of iterations is reached.

3.3 YOLOv3: An incremental improvement

YOLOv3 (You Only Look Once, Version 3) [7] is a real-time object detection algorithm that identifies specific objects in videos, live feeds, or images. It is the third version of YOLO [52]. 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 416x416 image (as in Fig. 3.3). 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.

The YOLOv3 algorithm first separates an image into a grid. Each grid cell predicts some number of boundary boxes (sometimes referred to as anchor boxes) around objects that score highly with the aforementioned predefined classes. Each boundary box has a respective confidence score of how accurate it assumes that prediction should be, and detects only one object per bounding box. The boundary boxes are generated by clustering the dimensions of the ground truth boxes from the original dataset to find the most common shapes and sizes. YOLOv3 is fast and accurate in terms of mean average precision (mAP) [53, 54, 55] and intersection over union (IOU) [53] values as well. It runs significantly faster than other detection methods with comparable performance. YOLOv3 increased the AP for small objects by 13.3, which is a massive advance from YOLOv2.

3.4 Attention Mechanism and Visual Attention

The attention mechanism is one of the most valuable breakthroughs in Deep Learning research in the last decade. A neural network is considered to be an effort to mimic human brain actions in a simplified manner. Attention Mechanism is also an attempt to implement the same action of selectively concentrating on a few relevant things, while ignoring others in deep neural networks. The attention mechanism

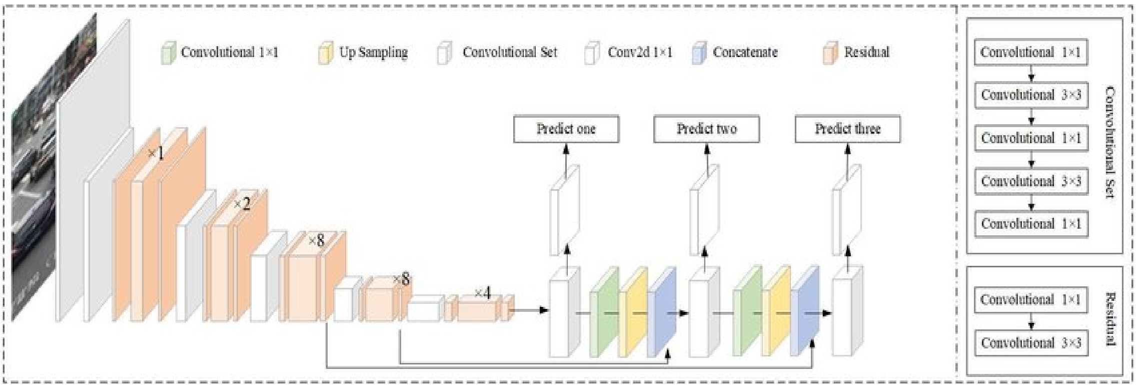

Figure 3.3: Structure detail of YOLOv3.It uses Darknet-53 as the backbone network and uses three scale predictions. Image from [56].

emerged as an improvement over the encoder decoder-based neural machine translation system in natural language processing (NLP). Later, this mechanism, or its variants, was used in other applications, including computer vision, speech processing, etc. In short, there are two RNNs [57] /LSTMs [58]. One it is called the encoder – this reads the input sentence and tries to make sense of it, before summarizing it. It passes the summary (context vector) to the decoder which translates the input sentence by just seeing it (Fig. 3.4). The main drawback of this approach is evident. If the encoder makes a bad summary, the translation will also be bad. And indeed it has been observed that the encoder creates a bad summary when it tries to understand longer sentences. It is called the long-range dependency problem of RNN/LSTMs. RNNs cannot remember longer sentences and sequences due to the vanishing/exploding gradient problem. It can remember the parts which it has just seen. Although an LSTM is supposed to capture the long-range dependency better than the RNN, it tends to become forgetful in specific cases. Another problem is that there is no way to give more importance to some of the input words compared to others while translating the sentence.

In addition to models that incorporate the concept of Attention Mechanism, there are also procedures that allow the standard deep learning models to focus on specific features. This allows for a more robust model.Attention can be applied to any kind of inputs, regardless of their shape. In the case of matrix-valued inputs, such as images, we can talk about visual attention. Let $I \in R^{H \times W}$ be an image and $g \in R^{h \times w}$ an attention glimpse i.e. the result of applying an attention mechanism to the image I.

An attention map is a scalar matrix representing the relative importance of layer activations at different 2D spatial locations with respect to the target task, i.e., an attention map is a grid of numbers that indicates what 2D locations are important for a task. Important locations correspond to bigger numbers and are usually depicted in red in a heat map. There are two different terms for the concept of attention [59]:

- Soft attention uses “soft shading” to focus on regions. Soft attention can be learned using good old backpropagation/gradient descent (the same methods that are used to learn the weights of a neural network model.) Soft attention maps typically contain decimals between 0 and 1.

- Hard attention uses image cropping to focus on regions. It cannot be trained using gradient descent because there’s no derivative for the procedure “crop the image here.” Techniques like REINFORCE [60] can be used to train hard attention mechanisms. Hard attention maps consistent entirely of 0 or 1, and nothing in-between; 1 corresponds to a pixel that is kept, and 0 corresponds to a pixel that is cropped out.

  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.[61] 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 [62] 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 [63] 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. [64] 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.

Attention mechanisms have been successfully applied in several contexts. The first part of [65] 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 [66] it is proposed a network composed of two main modules, namely a re-identification (Re-ID) module,

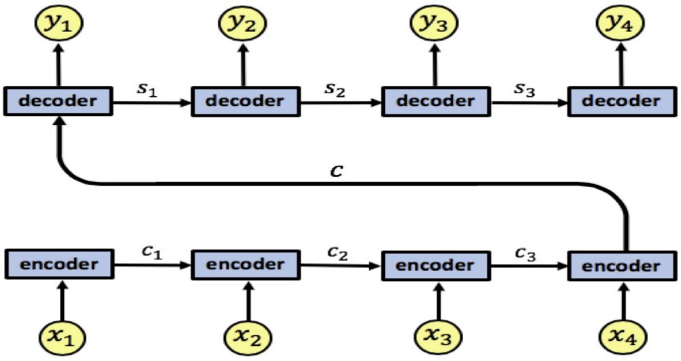

Figure 3.4: Simple representation of traditional Encoder-Decoder architecture using RNN/LSTM. Image from [68].

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 [67] 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.

3.5 Conditional generative adversarial nets

Introduced in 2014 by University of Montreal PhD student Mehdi Mirza and Flickr AI architect Simon Osindero, Conditional GAN[10] is a generative adversarial network whose Generator and Discriminator (Fig. 3.5) are conditioned during training by using some additional information. This auxiliary information could be, in theory, anything, such as a class label, a set of tags, or even a written description. Conditional generative adversarial network, or cGAN for short, is a type of GAN that involves the conditional generation of images by a generator model. GANs rely on a generator that learns to generate new images, and a discriminator that learns to distinguish synthetic images from real images. In cGANs, a conditional setting is applied, meaning that both the generator and discriminator are conditioned on some sort of auxiliary information (such as class labels or data) from other modalities. As a result, the ideal model can learn multi-modal mapping from inputs to outputs by being fed with different contextual information. Even the random distribution that the fake images follow will have some patter; it is possible to control the output of the generator at test time by giving the label for the image you want to generate. This type of network works in the following way:

1. The generator takes a random noise and a one-shot encoded class label as input. And it produces a false image of a particular class.

2. The discriminator takes an image with one-hot labels added as depth to the image (channels), i.e. for an image of 28 * 28 *1 size and a one-hot vector of size n, the image size will be 28 * 28 * (n+1).

3. The discriminator produces whether the image belongs to that class or not, i.e. real or false.

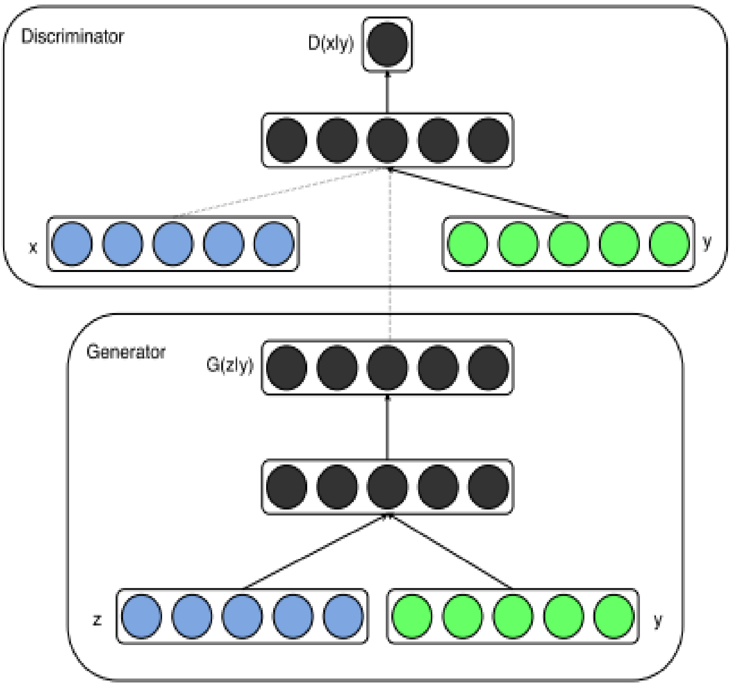

Figure 3.5: Conditional adversarial net. Image from [10].

3.6 Edgeconnect:Generative image inpainting with adversarial edge learning

In [69] the authors propose a new approach for image inpainting. This network includes an edge generator followed by an image completion network. The edge generator is capable of hallucinating edges in missing regions given edges and grayscale pixel intensities of the rest of the image; so it hallucinates the edges of the missing region (either regular or irregular) of the image. The image completion network combines edges in the missing regions with color and texture information of the rest of the image to fill the missing regions. They also propose an end-to-end trainable network that combines edge generation and image completion to fill in missing regions

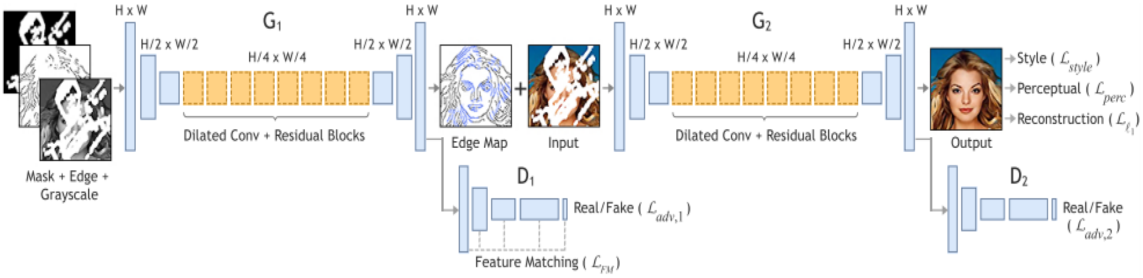
H x W
H/2 x W/2
G$_1$
H/4 x W/4
H/2 x W/2
H x W
Dilated Conv + Residual Blocks
Mask + Edge + Grayscale
Edge Map
Input
H x W
H/2 x W/2
G$_2$
H/4 x W/4
H/2 x W/2
H x W
Dilated Conv + Residual Blocks
Output
Style ($\mathcal{L}_{style}$)
Perceptual ($\mathcal{L}_{perc}$)
Reconstruction ($\mathcal{L}_{\ell_1}$)
D$_1$
Real/Fake ($\mathcal{L}_{adv,1}$)
Feature Matching ($\mathcal{L}_{FM}$)
D$_2$
Real/Fake ($\mathcal{L}_{adv,2}$)

Figure 3.6: Incomplete grayscale image and edge map, and mask are the inputs of G1 to predict the full edge map. Predicted edge map and incomplete color image are passed to G2 to perform the inpainting task. Image from [69].

showing fine details. Specifically, in their methodology includes: a first network called Edge Generator, the Inpaint Network and another called Image Completion Network. Both stages follow an adversarial model, i.e. each stage consists of a generator/discriminator pair (Fig. 3.6).

The generators consist of encoders that down-sample twice, followed by eight residual blocks [70] and decoders that upsample images back to the original size. Dilated convolutions with a dilation factor of two are used instead of regular convolutions in the residual layers, resulting in a receptive field of 205 at the final residual block. For discriminators, they use a 70×70 PatchGAN [71, 34] architecture. They also use instance normalization [72] acrossall layers of the network.

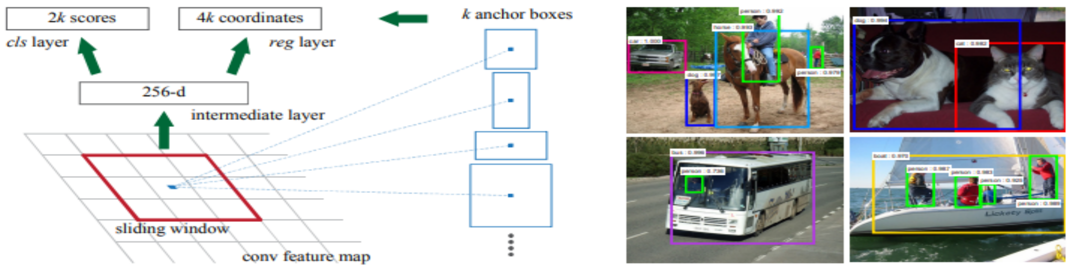

Figure 3.7: Region Proposal Network (RPN) and some example detections using RPN proposals on PASCAL VOC 2007 test (Image from [73]).

3.7 Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks (Faster RCNN)

One of the most popular object detection methods is the R-CNN series, developed by Ross Girshick et al in 2014, improved upon with Fast R-CNN [74] and then finally with Faster R-CNN [73]. The differentiating approach that makes Faster R-CNN better and faster is the introduction of Region Proposal Network (RPN). RPN is a fully convolutional network, trained end-to-end, that simultaneously predicts object boundaries and object scores at each detection.

A Faster R-CNN object detection network [73] is composed of a feature extraction network which is typically a pretrained CNN, similar to what we had used for its predecessor.

This is then followed by two subnetworks which are trainable. The first is a Region Proposal Network (RPN), which is, as its name suggests, used to generate object proposals and the second is used to predict the actual class of the object. So the primary differentiator for Faster R-CNN is the RPN which is inserted after the last convolutional layer. This is trained to produce region proposals directly without the need for any external mechanism like Selective Search. After this we use ROI pooling and an upstream classifier and bounding box regressor similar to Fast R-CNN [74]. As the feature extraction, ROI pooling and classifier are the same as the previous versions. The goal of RPN is to output a set of proposals, each of which has a score of its probability of being an object and also the class/label of the object. RPN can take any sized input to achieve this task. These proposals are further refined by feeding to 2 sibling fully connected layers-one for bounding box regression and the other for box classification i.e is the object foreground or background. The RPN that generates the proposals slide a small network over the output of the last layer of the feature map. This network uses an nxn spatial window as input from the feature map. Each sliding window is mapped to a lower dimensional feature.The position of the sliding window provides localization information with reference to the image while the regression provides finer localization information. Anchor boxes are some of the most important concepts in Faster R-CNN. These are responsible for providing a predefined set of bounding boxes of different sizes and ratios that are used for reference when first predicting object locations for the RPN. The original implementation uses 3 scales and 3 aspect ratios, which means k=9. If the final feature map from feature extraction layer has width W and height H , then the total number of anchors generated will be W*H*k.Anchor boxes at each spatial location, mark an object as foreground or background depending on its IOU threshold with the ground truth. All the anchors are placed in a mini-batch and trained using softmax cross entropy to learn the classification loss and smooth L1 loss for regression. NMS is the second stage of filtering used to get rid of overlapping boxes. In Figure 3.7 the architecture of Region Proposal Network (RPN) and some example detections using RPN proposals on PASCAL VOC 2007 test.

3.8 Focal loss for dense object detection (RetinaNet)

Lin et al. [75] highlight that the one-stage detectors that are applied over a regular, dense sampling of possible object locations have the potential to be faster and simpler, but have trailed the accuracy of two-stage detectors thus far. So they investigate why this is the case. They discover that the extreme foreground-background class imbalance encountered during training of dense detectors is the central cause. They introduce to address this class imbalance by reshaping the standard cross entropy loss such that it down-weights the loss assigned to well-classified examples. Their novel Focal Loss focuses training on a sparse set of hard examples and prevents the vast number of easy negatives from overwhelming the detector during training. They design and train a simple dense detector that call RetinaNet which uses a Feature Pyramid Network(FPN) [76] with Resnet [70] Backbone. FPN involves adding top level feature maps with the feature maps below them before making predictions. Adding a top level feature map with a feature map below usually involves upscaling the top level map, dimensionality matching of the map below using a 1x1 conv and performing element wise addition of both. One more important aspect is the initialization of model probabilities for foreground class before start of training. All positive anchors are assigned a prior probability of 0.01 so that they contribute more to the loss and to make sure large number of negative examples do not hamper training during the initial stage.

There are four major components of a RetinaNet model architecture (Fig. 3.8):

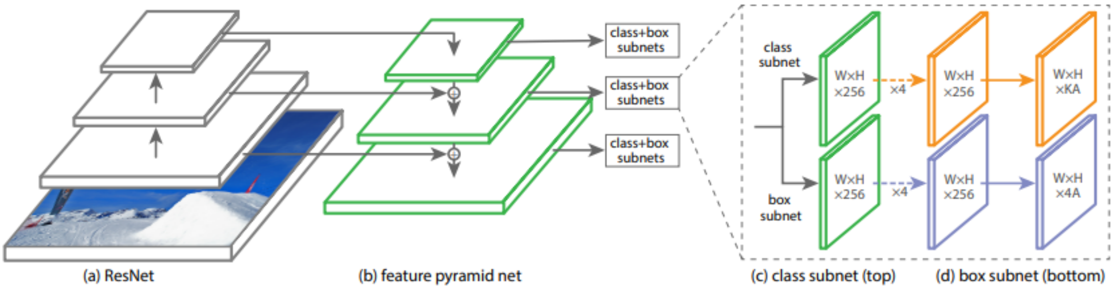

Figure 3.8: The one-stage RetinaNet network architecture uses a Feature Pyramid Network (FPN) backbone on top of a feedforward ResNet architecture (a) to generate a rich, multi-scale convolutional feature pyramid (b). To this backbone RetinaNet attaches two subnetworks, one for classifying anchor boxes (c) and one for regressing from anchor boxes to ground-truth object boxes (d). Image from [75].

- Bottom-up Pathway - The backbone network (e.g. ResNet) which calculates the feature maps at different scales, irrespective of the input image size or the backbone.

- Top-down pathway and Lateral connections - The top down pathway upsamples the spatially coarser feature maps from higher pyramid levels, and the lateral connections merge the top-down layers and the bottom-up layers with the same spatial size.

- Classification subnetwork - It predicts the probability of an object being present at each spatial location for each anchor box and object class.

- Regression subnetwork - It's regresses the offset for the bounding boxes from the anchor boxes for each ground-truth object.

3.9 EfficientDet: Scalable and Efficient Object Detection

EfficientDet [77] is an object detector which achieves state-of-the-art accuracy while being up to 9x smaller and using significantly less computation compared to prior state-of-the-art detectors. It's based on the traditional idea of running the algorithm on multiple resolutions of the same image hoping to capture both small and large scale phenomena. EfficientDet detectors are single-shot detectors much like SSD [78] and RetinaNet. Here, however instead of using the image at different resolutions authors use feature maps at different resolutions. The paper makes two major modifications to EfficientNet, namely BiFPN (or Weighted Bi-directional Feature Pyramid Network) and a new compound scaling method. EfficientDet uses the same backbone as EfficientNet but adds a bi directional feature pyramid network to help in multi scale feature fusion. BiFPN has 5 modifications over a normal FPN (Fig. 3.9): (1) Instead of only top-down feature, it adds another bottom-up feature fusion branch. (2) It has skip connections from the initial feature map to the fused feature map. (3)Nodes with only one input are removed, cause they do not do much fusion as other nodes. (4) The entire module is repeated multiple times. (5) Features are not summed directly, instead a weighted average is used hoping different resolution feature maps contribute to the fusion at different capacity. Unbounded weights bring problems in backprop, so we need to normalise it. They tried applying softmax to the weight values which worked but slowed down training. So a simple average after relu activation is used to normalise the weights. The need for a new scaling technique comes from the fact that there is the BiFPN as an additional module in the network and that too can be scaled. But there's no heuristic given

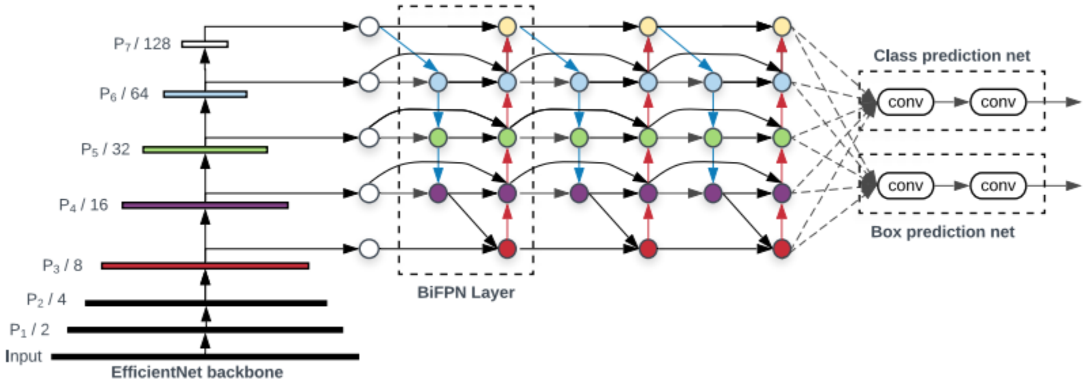

Figure 3.9: EfficientDet architecture. EfficientDet uses EfficientNet as the backbone network and a newly proposed BiFPN feature network. Image from [80].

about the scaling technique. The backbone networks are ImageNet[79] pretrained EfficientNets[80].

Chapter 4

Hierarchical Fine Tuning and Clustering

Ideally, a benchmark dataset should contains thousand of images related to a very high number of different polyps. A large dataset is useful also to properly train polyps detectors based on deep learning architectures. Despite the advancement ofthe state-of-the-art in the context of objects detection, methods suffer in recognising small polyps, which are the most importantto detect since are the one appearing at the beginning of thelesion progression.

We point out that organizing the training of a polyp detector in a hierarchical way in order to consider variabilities related to the visual content of the colonoscopy images, as well as the one related to the size of polyps, help to improve the detection performances by reducing the false negatives (i.e, the number of lesions not detected) especially for small size lesions (i.e., the most important to detect especially in early stage of the colon cancer). To perform our study, we have collected and labelled a novel real large-scale dataset which is at least twenty times larger than the datasets currently available in literature. To demonstrate that a hierarchical organization of the training procedure can help to improve the results of an object detector, we have used the popular YOLOv3 detector [52, 7] as baseline.

Experiments point out that, considering the proposed hierarchical organization of training, polyps can be detected with a per-polyp precision (recall) of 99.46% (92.03%) and a per-frame precision (recall) of 95.06% (66.95%).

4.1 Methods

A standard way to proceed in building a polyp detector is to choose a deep learning architecture and train it on a labeled dataset to regress the bounding box coordinates related to the polyps belonging to the image under consideration. Considering that the tissues surrounding polyps (i.e., normal mucosa) have a huge variability in appearance making difficult the detection of polyps especially when very small and between folds, as first variant we considered to fine-tune the polyp detector after partitioning the dataset with respect to visual content. In the first stage a detector D is trained on the overall training set until convergence. After this first training, the backbone of the detector can be used as method to extract feature by removing the last layers of the network dedicated to the regression of the bounding box. In this way, all the training samples can been represented with a feature vectors containing information about the visual content. This new input space can be clustered to obtain K modalities and considering the original image samples belonging to each modality the detector D can be fine-tuned to obtain K new detectors D_1, ..., D_K specialised on each modality. This procedure of training is hence composed by a hierarchy of one level.

Another hierarchical way to specialise the detector is to consider the known size of the polyps belonging to the training set. Also in this case, the samples of the training set can be partitioned in groups containing small, medium and large polyps. This can be done by simply asking an expert to choose two threshold which define the three different size. Hence, a second hierarchy with one level can be obtained considering the detector D trained on the overall dataset and specialising it by fine-tuning with respect to the three groups. This procedure will give three new detectors D_s, D_m, D_l.

The third hierarchical way we have considered to fine-tune the object detector is the one in which both modalities visual content and size of polyps are exploited. To this aim after training a detector D we have performed a clustering on the obtained visual features to obtain K groups. Each group of samples belonging to each modality have been further partitioned considering the dimension of polyps. Hence this procedure produce $K * 3$ detectors $D_{1,s}$, $D_{1,m}$, $D_{1,l}$,..., $D_{K,s}$, $D_{K,m}$, $D_{K,l}$ in a two level hierarchy.

In sum, starting from a detector D trained on the overall dataset, we have considered three possible hierarchical variants:

1. a one level hierarchy where the detector D is trained on the overal dataset and then fine-tuned considering different appearance modalities;

2. a one level hierarchy where the detector D is further fine-tuned considering the dimension of polyps;

3. a two level hierarchy where the detector D is fine-tuned considering both appearance and size variabilities.

To perform inference after fine-tuning with one of the considered hierarchies, a new pattern is feed to each specialised network and the results are combined to obtain the final detection (i.e., the bounding box of all networks are considered).

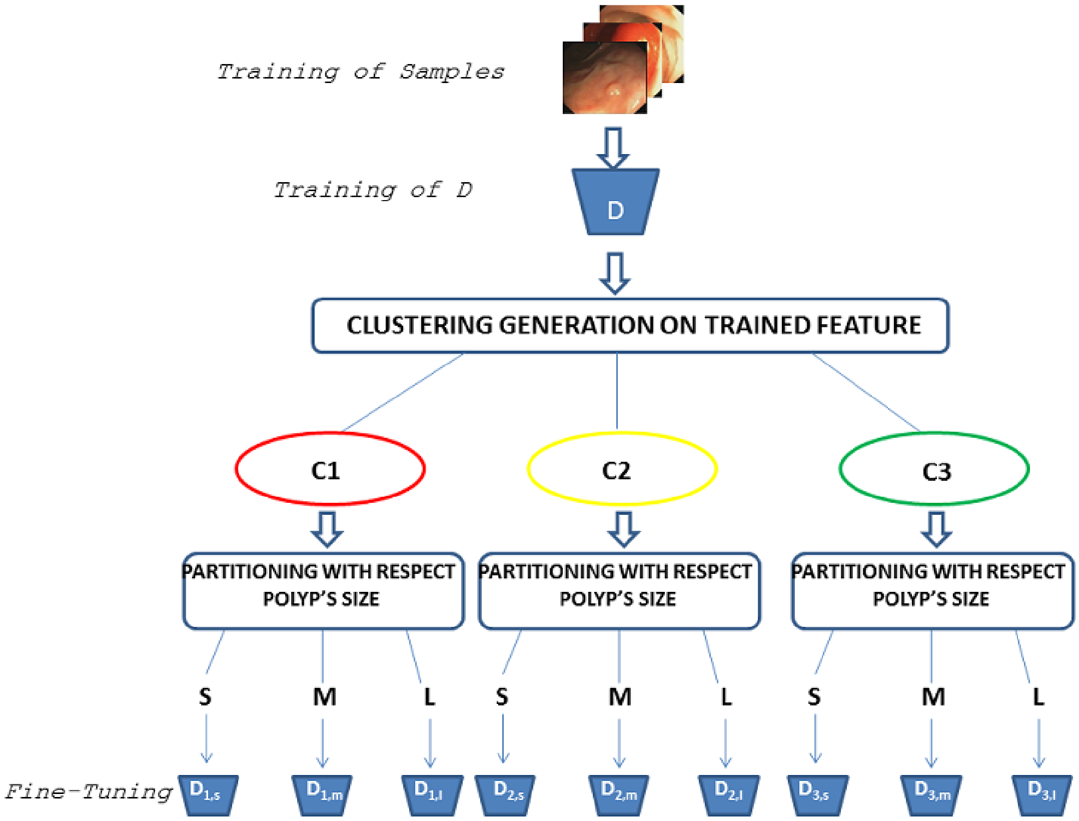
Training of Samples
Training of D
D
CLUSTERING GENERATION ON TRAINED FEATURE
C1
C2
C3
PARTITIONING WITH RESPECT POLYP'S SIZE
PARTITIONING WITH RESPECT POLYP'S SIZE
PARTITIONING WITH RESPECT POLYP'S SIZE
S
M
L
S
M
L
S
M
L
Fine-Tuning
$D_{1,s}$
$D_{1,m}$
$D_{1,l}$
$D_{2,s}$
$D_{2,m}$
$D_{2,l}$
$D_{3,s}$
$D_{3,m}$
$D_{3,l}$

Figure 4.1: Training phase for the two level hierarchy.

Figures 4.1 and 4.2 illustrate the two level hierarchy considered in our experiments for training and test phases. As baseline detector D we have used the popular YOLOv3 [7], whereas the K-means clustering has been employed to built the modalities related to the features extracted considering visual content. Among the value of K which have been considered, we report results obtained with $K = 3$ because have obtained the best performances. The three groups corresponding to the modality related to the size of polyps have been produced considering the feedback of colonoscopy experts.

Table 4.1: Cross-Dataset evaluation comparison.

Train Set	**Test Set**	**Prec**	**Rec**	**F_1**
CVC-612	CVC-612	83.22	59.90	69.66
CVC-12k	CVC-612	92.62	66.67	77.53
Our Dataset	**CVC-612**	**99.35**	**74.40**	**85.08**
CVC-612	CVC-12k	71.78	27.51	39.77
CVC-12k	CVC-12k	86.64	31.50	46.21
Our Dataset	**CVC-12k**	**94.46**	**74.76**	**83.46**
CVC-612	Our Dataset	21.16	02.99	05.24
CVC-12k	Our Dataset	52.93	15.82	24.36
Our Dataset	**Our Dataset**	**95.02**	**64.53**	**76.86**

4.2 Experimental Settings and Results

One of the main problem in assessing the performances of polyp detection framework is the lack of publicly available and representative dataset to perform experiments. The most popular dataset available is the CVC-12k [41, 42]. It is composed by only 11,954 images with ground truth masks related to only 18 different videos (i.e., 18 patients). Most important, images of this dataset are very redundant (more 10K images from only 18 videos) and have small resolution (384x288 pixels), which means that some important features that could be useful during training to set parameters of a deep learning based detector to distinguish polyps tissues from normal mucosa could have been destroyed in the resizing. Other dataset in literature in this context, such as CVC-356 and CVC-612 [27] are much smaller than CVC-12k and have lesions which are clearly visible and mostly of them are centered in the frames. In all the previous mentioned datasets there is low variability in terms of polyps type (i.e., flat vs others) and few polyps between folds.

Due to the above limitations, for the experiments of this paper we have created a novel dataset twenty times larger than CVC-12k (i.e., $> 200K$ images), with images

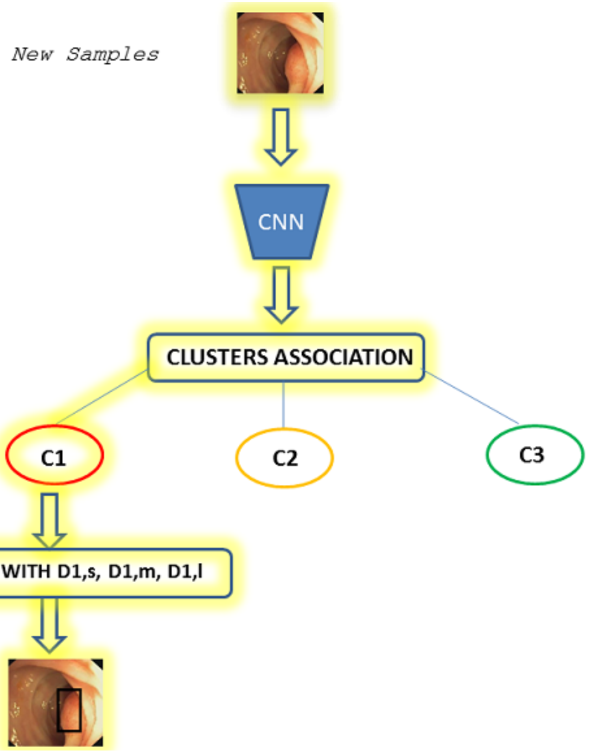

Figure 4.2: Test phase for the two level hierarchy.

sampled and labelled by colonoscopy experts considering more than 180 videos. In the considered dataset are represented more than 500 different polyps of different type (e.g. flat vs raised polyps), size, positions in the image and also presenting occlusions (polyps between folds).

To evaluate the quality of our dataset we have performed a cross-dataset polyp detection evaluation. Specifically, we have partitioned randomly our dataset, as well as the CVC-612 and CVC-12k datasets obtaining for each of them training (70%) and test (30%) sets[1]. We have then trained YOLOv3 with the training set obtained from one of the three considered datasets (e.g., CVC-12k) and we have tested it on the test sets related to all the datasets (i.e, CVC-12k, CVC-612 and our dataset).

The results with respect to these nine cross-dataset experiments are reported in

[1] CVC-356 has been excluded by this test because it is too small.

Table 4.1 considering classic evaluation measures used in the field on per-frame detection bases: prevision (Prec), recall (Rec), F_1 score. The best results with respect to each test set are reported in bold, whereas second best results are underlined. It is clear that the proposed dataset is more challenging than CVC-12K (compare precision and recall results obtained training with our dataset and testing on all the other datasets) and that already training with our dataset, rather than CVC-12K, and testing on CVC-12K it is possible to obtain an improvement on overall measures with a big margin (more than 7% in precision and more than 40% on recall - see rows five and six in Table 4.1). There is also a clear evidence that an appropriate and realistic dataset can help to better train a deep learning based detector. Indeed, in all cases, the best results are obtained training the considered baseline detector with our dataset. Specifically, comparing with respect to the second best results (i.e., the once obtained training with CVC-12K) there is an improvement of the precision of about 7% on CVC-612 test set, of about 7% on the CVC-12K test set, and of about 42% on our test set. Recall improves as well of about 8%, 43%, and 49% considering the CVC-612, CVC-12K and our test sets respectively. It is worth noting that the performances improve with a considerable margin also with respect to F_1 score in all cases. For example, when the test set related to our dataset is considered in the experiments, there is an improvement of more than 50% for the F_1 score with respect to the second best. Taking into account this, we have continued our experimentation only considering our dataset.

Each further experiment involving our dataset has been repeated three times on three fixed randomly obtained partitions forming training and test sets. Experiments have been evaluated with the aforementioned measures considering both per-frame and per-polyp detection. Indeed, despite the per-frame measures consider

Table 4.2: Per-frame comparison.

Hierarchies	**Prec**	**Rec**	$\mathbf{F}_1$
D	95.02	64.53	76.86
$D_1, ..., D_K$	93.75	65.83	77.34
D_s, D_m, D_l	95.06	66.95	78.56
$D_{1,s}, D_{1,m}, D_{1,l}, ..., D_{K,s}, D_{K,m}, D_{K,l}$	94.99	66.60	78.34

the detections on per-frame bases, per-polyps measures consider the case of polyps which are detected at least in one frame of the colonoscopy. The per-polyp measures are also important because in real clinical systems it is fundamental to perform the detection of each polyp. This means that polyps have to be detected at least in one frame in which they appear during a colonoscopy screening. This is very relevant in clinical systems in order to notify polyps to the experts at least in one frame where they appear in order they can perform a more accurate check. Differently, the per-frame measure do not consider the case in which a polyp is missed in all the frames of the colonoscopy.

The results of each experiment have been obtained by averaging the results over the three partitions of the dataset. We have performed each training-test partition such that the distribution over the size of the polyp is as similar in both training and test. To do so, each training-test split has been obtained performing 100 random split of our videos in two separate subset to guarantee that an image of a video belong only to the training or the test set. Each of these 100 random training-test split has been performed in order to obtain a training set with a number of images of about 70% and a test set of about 30%. For each of the 100 random training-test split we have computed the distribution of the size of the polyps for both training and test sets and the histogram intersection between each of the two obtained distributions and the distribution of the overall dataset. The intersection values have been summed and stored as similarity score for each random training-test split of videos. The final training and test sets used for each experiments have been hence obtained considering the random spit with the highest similarity score. This splitting procedure was done to ensure that images of a video are considered only for training or testing purpose and to be sure that both training and testing sets contain high variability with respect to the size of the lesions. The results obtained with the different hierarchical variants of the considered object detector (see previous section) are reported in Table 4.2 and Table 4.3 considering per-frame and per-polyps evaluation respectively.

Results on the per-frame basis point out that in all cases the hierarchies help to improve the recall measure with respect to the baseline detector D, whereas the precision has no significant drop. Also F_1 measure is in favour of this. It is worth noting that exploiting one or two level hierarchy an improvement is always observed considering the per-polyps basis results. This means that the hierarchical frameworks were able to detect polyps missed from the baseline detector D, and this has an important clinical impact. In general results point out that the one layer hierarchy considering size is to be preferred and that a combination of visual and size modalities in a two layer hierarchy does not further improve the results. Last but not least, a qualitative assessment of the polyps detected in the hierarchical frameworks has pointed out that the recovered detections (which are missed in the baseline) are related to small polyps, which usually are the one at the early stage of the disease and very difficult to be detected.

Table 4.3: Per-polyps comparison.

Hierarchies	**Prec**	**Rec**	$\mathbf{F}_1$
D	97.84	91.91	94.78
$D_1, ..., D_K$	98.40	92.96	95.60
D_s, D_m, D_l	99.46	92.03	95.60
$D_{1,s}, D_{1,m}, D_{1,l}, ..., D_{K,s}, D_{K,m}, D_{K,l}$	99.46	92.03	95.60

4.3 Discussion

This chapter considered the problem of polyps detection in colonoscopy images. The main problem in this domain is the identification of polyps at their early stage, i.e., polyps with very small size. To improve the ability of a polyps detector we have considered to look at the visual and size variabilities by performing a hierarchical training procedure. The experiments on a novel real large dataset confirm that a hierarchical framework can improve the results and more specifically can obtain better performances on lesions of small size.

Chapter 5

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

In the previous study we proposed a model to detect a lesion in a static image, namely no temporal information were considered in the decision process. Differently on our previous work, here we explore the possibility to exploit temporal redundancy to improve the detection performance. 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. 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. 5.1). Experiments, conducted using a modified version of YOLOv3 [7] 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. 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 of the polyps, which cause confusion in detection and fall into false negatives. 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. In summary, 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.

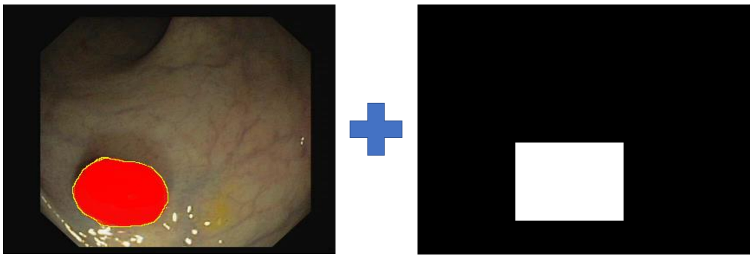

Figure 5.1: Attention Mechanism. The procedure for creating input images of each video sequence by the combination of the original frame (left) and Mask related to the previous frame (right).

5.1 Methods

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 (with the coordinates of the bounding box enclosing the position of the polyp in the frame) of the polyps in the previous frame when this last information is available. We then exploit the coordinates of the bounding-box enclosing the position of the polyp in the previous frame.

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. We also show results demonstrating improved performance by attending to false negatives, which are caused by polyps that are located at the edges of the frame or confused with the probe light or confused with the mucosa due to the very similar texture.

5.1.1 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 the bounding box annotation of the frame F_j (Fig. 5.2). 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 [7] 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.

5.1.2 The YOLOv3

In this work we choose YOLOv3 architecture as polyps detector, which is the third version of YOLO [52]. It presents better backbone classifier with respect to the

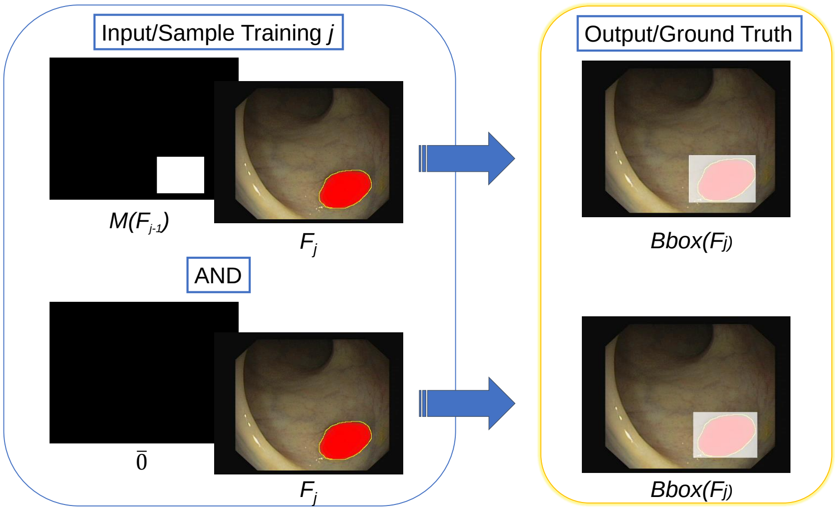

Figure 5.2: Construction of the input for the training. Given a frame F_j, it is combined with the binary mask $M(F_{j-1})$. The same Fj is also combined with the mask $\bar{0}$ in order to train the network even when no polyps occurs in the previous frame. The input data of the object detector is a four-channel signal.

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 × 416 × 3 tensor. However, we change the input layer to make YOLOv3 able to accept 416 × 416 × 4 tensor, in which the new channel includes the binary attention mask. More details can be found in 3.3.

5.2 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.

5.2.1 Dataset

We used our dataset [43]. Our contribution is based on exploiting the information of colonoscopy video sequences, which are nothing more than temporal frames.
The dataset has a high variability in terms of size of polyps (Fig. 5.4). Our idea is not applicable with the datasets available in literature [27], [41], [42] 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.
More details about our dataset is found in Section 2.4.

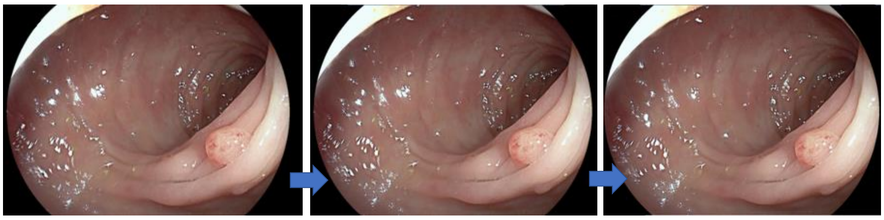

Figure 5.3: Example of our video sequence.

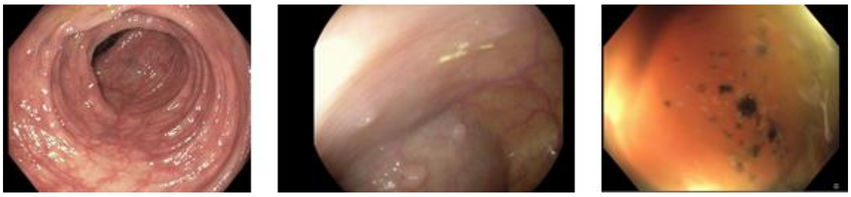

Figure 5.4: Variability of the polyps and mucosa in the dataset.

5.2.2 Evaluation Metrics

Performances are evaluated by popular metrics: Precision ($Prec$), Recall (Rec) and F1-score ($F1$) [81]. 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.

5.2.3 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. 5.5 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.

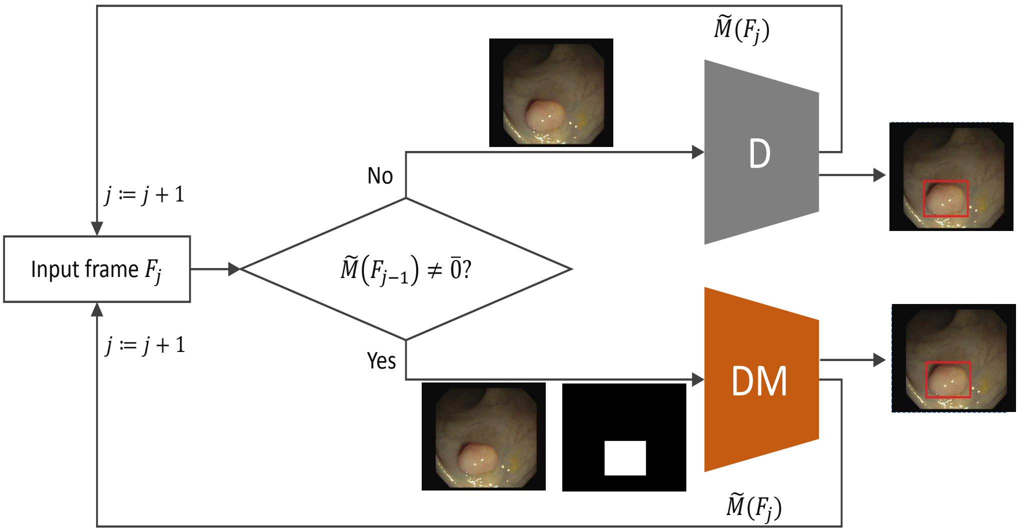

Figure 5.5: Combination of the two detectors. The sample F_j is given to the system and the attention mask $\tilde{M}(F_{j-1})$ is checked. If it is available $(\tilde{M}(F_{j-1}) = \bar{0})$, F_j and the related mask are concatenated and input into the model DM. Else, if no attention mask is available, only the frame F_j is fed to the standard model D.

5.3 Results

In this section we discuss experimental results achieved with the proposed framework. In Table 5.1 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 improvement of 4.63%, which

Table 5.1: Evaluation comparison

Detector	**Prec**	**Rec**	$\mathbf{F_1}$
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

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.

5.4 Discussion

In this chapter, we have proposed 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.

Chapter 6

Inpainting Based Data Augmentationto Improve Polyp Detection in Colonoscopy

In this chapter we present a further investigation to improve polyp detection performance.

The need to have large, variable and well-labeled dataset led the researchers in the field to develop data augmentation approach to perform a proper training by avoiding both underfitting and overfitting. The most common and naive data augmentation procedure for images, consists of applying different affine transformations to have the same frame rotated, scaled, translated, or flipped. However, this kind of augmentation slightly affect the training performance since just introduces a minor geometric variability of the point of a polyp. A proper synthetic data generator should mimic the original data samples by introducing a higher variability and not just increasing the data samples numerosity. To this aim, deep learning have recently been employed in synthetic image generation. Specifically, Generative Adversarial Networks (GANs) [9] and conditional GANs [10] have been adopted to address the problem of smart augmentation approaches and, recently, many medical imaging analyses and applications have been reported to successfully use GAN. Such applications include medical image segmentation, classification, inpainting, and so on. Also, for automatic polyps detection, recent literature works have adopted conditional GANs by exploiting the mucosa and polyp edges to generate more realistic synthetic images [82]. Inspired by such works, we address the problem of data augmentation for colonoscopy dataset in order to train better detection models than the ones obtained by only using the original annotated data.

However, differently from previous researches [82], we propose to exploit inpainting to generate synthetic polyps images through a deep neural network [69]. Specifically, we exploit the original label masks and also the boundary between the polyps and the healthy mucosa to train an augmentation network able to generate new photorealistic lesions. Hence, we improve polyps detection performance by training a detector with an augmented datasets obtained with the proposed augmentation framework. We prove that the employed inpainting network allows to smartly augment the polyps images in order to guarantee an improvement in the detectors performance thanks to a more effective training phase. In order to evaluate our framework, we adopt different publicly available datasets and we also perform cross-dataset evaluation by training the detector using datasets different from the ones employed for testing. We also prove the validity of our proposal by evaluating it using different state-of-art detectors.

6.1 Proposed Method

We introduce the literature methods and data we employed in our proposal, i.e. the inpainting network [69], the datasets and the detectors using for testing. Secondly, we describe how we use the inpainting network for synthetic polyp images generation.

6.1.1 Edgeconnect for Data Generation

Nazeri et al. [69] propose a new approach for image inpainting, which includes an edge generator followed by an image completion network (Fig. 6.1).

Their model is trained on the irregular mask dataset, while to test the model, they test separately on the edge model, then the inpaint model, and finally the joint model. The mask shape covers the entire region of the mask in the input image. Although they point out, however, that the models can be merged with a joint model, they also say that it is possible to train the parts separately and use them that way. They propose such neural network for image editing purposes, such as object removal and scene generation. To train the inpainting Network generator, they generate the training labels (i.e., edge maps) using the Canny edge detector[83]. The sensitivity of the Canny edge detector is controlled by the standard deviation of the Gaussian smoothing filter. For more details, view the Section 3.6.

6.1.2 Detectors

Here we report a short description of Faster R-CNN [73], RetinaNet[75] and the EfficientDet[77] that we use as detectors to validate our contribution.

Faster R-CNN. This architecture is composed of two modules: a region proposal network (RPN) to select candidate object regions and the detector Fast R-CNN

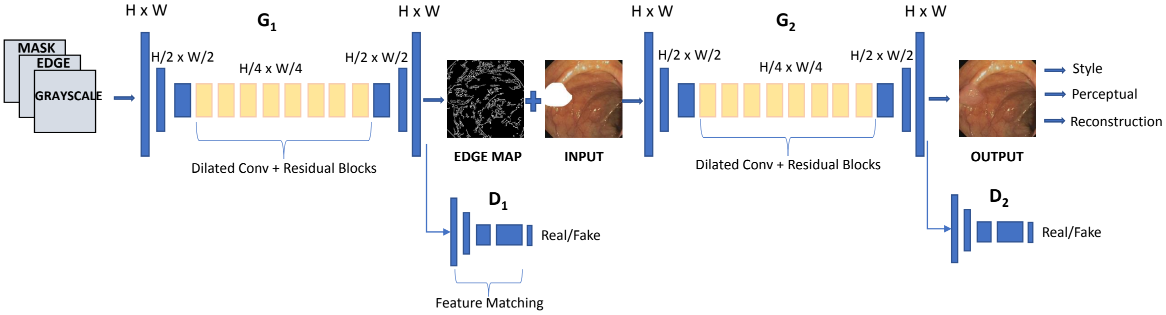

Figure 6.1: EdgeConnect proposed by [69].

[84]. RPN acts as attention mechanism module which tells the Fast R-CNN where to look. This network region proposal step is nearly cost-free and enables a unified, deep-learning-based object detection system to run at near real-time frame rates. Furthermore, the RPN improves region proposal quality, which drives a general increasing in object detection accuracy. More details can be found in Section 3.7.

RetinaNet. Another detector used to validate experiments is RetinaNet. It is a one-stage object detection model that utilizes a focal loss function. RetinaNet improves performance by additionally using the Feature Pyramid Network (FPN) [76] for feature extraction. The backbone is responsible for computing a convolutional feature map over an entire input image. The first subnet performs convolutional object classification on the backbone's output; the second subnet performs convolutional bounding box regression. FPN concatenates feature maps from layers at different depths to improve detection at each scale. Another relevant aspect of this model is the use of focal loss to solve the class imbalance problem, that is the imbalance between background and foreground. Therefore, focal loss is introduced to assign higher weights to difficult foreground objects and lower weights to the easy background case. RetinaNet uses a dynamically scaled cross entropy loss where the scaling factor decays to zero when confidence in the correct class increases. Intuitively, this scaling factor can automatically down-weight the contribution of easy examples during training and rapidly led the model to focus on hard examples. More details of architecture is reported in Section 3.8.

EfficientDet. EfficientDet is a type of object detection model that uses a variety of backbone optimizations and tweaks, such as the bi-directional feature pyramid network (BiFPN), which allows easy and fast multiscale features fusion. Moreover, this model uses a compound scaling method that uniformly scales resolution, depth, and width for all the backbones, features network, and box/class prediction networks at the same time. EfficientDet detectors are single-shot detectors much like Single Shot Detector (SSD) and RetinaNet. The proposed BiFPN serves as the feature network which takes level 3–7 features P3, P4, P5, P6, P7 from the backbone network and repeatedly applies top-down and bottom-up bidirectional features fusion. More detailed information is given in the Section 3.9.

6.1.3 Datasets

The dataset used are the CVC-CLINIC or CVC-612 [27] and the CVC-ClinicVideoDB or CVC-12k [28] dataset (Fig. 6.2). The CVC-CLINIC dataset includes 612 polyp image frames with a pixel resolution of 388 × 284 pixels in SD (standard definition).

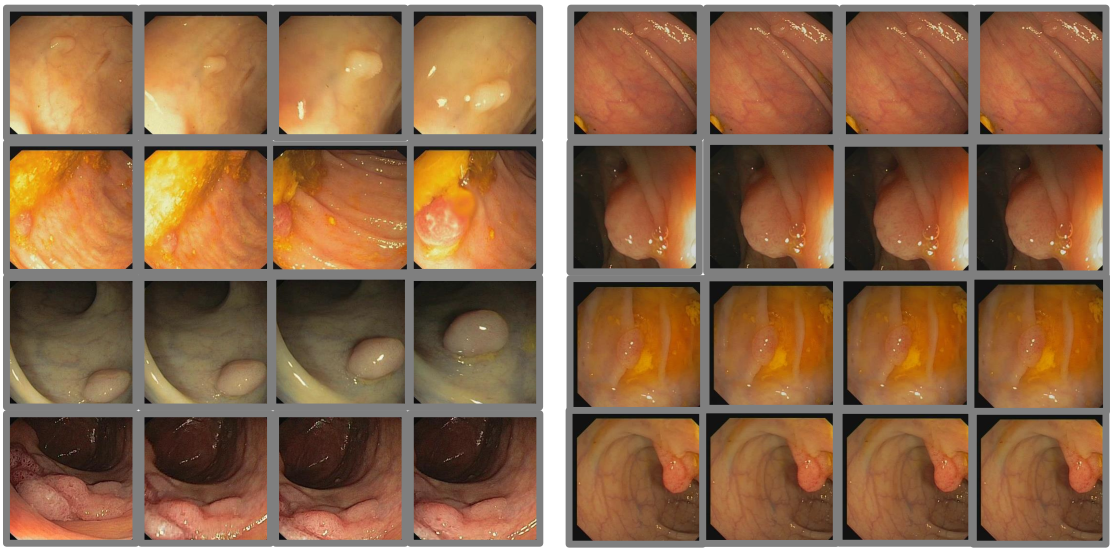

Figure 6.2: Four frame sequences from the CNC-CLINIC dataset (Left); Four frame sequences from the CVC-ClinicVideoDB dataset (Right).

The CVC-ClinicVideoDB video dataset comprises of 18 different SD videos of different polyps. In this dataset, 10025 frames out of 11954 frames contain a polyp, and the size of the frames is 384 × 288. Each frame in the video databases comes with a binary ground truth, in which each polyp is annotated by clinical experts. Each positive video includes a unique polyp.

According to COCO format, the annotations (i.e., polyps bounding boxes) are described in a single JSON file for the whole dataset. Object classes, just polyps in our study, are listed separately in the categories tag and identified by an ID. More details is given in Section 2.4.

6.1.4 Our Proposal

In our proposal, we use EdgeConnect to generate new images which depict synthetic polyps, while in the original work [69] the authors use it to fill generic missing regions and test it on standard datasets such as CelebA [85], Places2 [86] and Paris

StreeView [87]. Although the original inpainting network is composed of two parts, i.e. Edge model model and Inpaint model, in the present work we only train the second one. Generally, the Edge model is used to complete the edge map in images with missing parts to be inpainted. However, since we employ masks and edge maps of real polyps even at inference time, original fully edge maps are always available and we do not need to estimate such maps. Our idea is to train the Inpaint network only by using polyp images and by providing, as missing part to be inpainted, the area where a polyp is located. Hence, the Inpaint network learns to fill the missing regions always with a polyp, which makes it a synthetic polyp images generator.

The CVC-CLINIC dataset is used for training the inpainting network in order to build the proposed generator. The dataset is split into 80% for training, 10% for validation, and 10% for test and geometric data augmentation is employed to reduce overfitting in training phase. In particular, we apply horizontal and vertical flipping, 90°, 180°, and 270° rotation, 10% zoom-in. By combining all the transformations, we get 15 altered frames from the original one. Then we have 7648 frames after transformations for training, 1088 images for validation, and 1056 of test. To train the Inpaint module of EdgeConnect we need to extract edge maps from input frames; to this aim, we use Canny edge detector in accordance with [69]. It is applied with a high threshold 100 and low threshold equal to 20. Then, all the image frames are cropped to 256 × 256 and black borders are removed. After the Inpaint module of the EdgeConnect is properly trained, it can be used for inference, namely to generate new polyps image. To this aim, the new inpainting model has to be input with an RGB image, which presents a missing area located where a polyp has to be generated, and its edge map. Even if not mandatory, our suggestion is to synthesize new polyp images by using mask and edge of real polyps in order to get more realistic results.

6.2 Experimental Setting and Results

We evaluate the polyp detection performance by comparing the model trained with and without the proposed augmentation strategy.

6.2.1 Inference on EdgeConnect

CVC-ClinicVideoDB dataset includes many empty frames. For the generation of new frames with polyps we exploited the no-polyp frames. We selected those of good quality, that is, those that visually are better and without e.g. blurred effect, without water bubbles or without too much light reflected on the mucosa. From this visual analysis we selected 500 frames with mucosa, which were then used as base frames on which to generate polyps with our methodology. In order to increased the variability and get more realist images, we generate the synthetic dataset by randomly select the masks related to the polyp location from the validation and test set of CVC-CLINIC, where it was previously applied the classic data augmentation. These masks perfectly fit the shape of the polyp. For this reason, we decided to use masks from another dataset such as the CVC-CLINIC rather than the CVC-CLINICVideoDB because in the latter the masks have an ellipsoidal shape which grossly fit the lesions. The mask can have any shape, can be in any position, and in output we have a completely new frame.

For convenience, we refers the CVC-CLINICVideoDB dataset as CVCdb and the CVC-CLINIC as CVC. Once the best no-polyp frames were selected from the CVCdb dataset, these frames were cropped back to a size of 256×256. At the same time, the

Validation Set and Test Set frames used in the CVC dataset were taken for training the inpainting network, i.e. EdgeConnect. Of the latter, the relative mask and the creation of an Edge Map are fundamental for this phase, but only relative to the lesion. So the edge map, in this case, is all black, but with white edges corresponding only to the lesion. Randomly, one of these masks is selected to attach to the RGB image of the empty frame. Similarly, the edge map of the randomly selected polyp and the mucosal edge map of the original RGB image are combined (Fig. 6.3).

In order to have more variability in generating polyps on combining masks and no-polyp frames, we decide to distinguish between three kind of polyps in respect of their size: small, medium and large. The threshold which characterize the three size are the same used in [43]. This insight come from the observation that small polyps are, not surprisingly, more difficult to detect. Hence we prefer to generate more synthetic images with small polyps, specifically, 80% of small polyps images and 20% of medium polyps images.

6.2.2 Detection Evaluation Metrics

For a proper evaluation of the detector performance with and without data augmentation we employ metrics which are designed on average precision (AP) and average recall (AR). AP is a metric which depends on the area under the Precision versus Recall. It summarizes the precision-recall trade-off dictated by confidence levels of the predicted bounding boxes [53].

$$AP = \sum_n (R_n - R_{n-1}) P_n$$

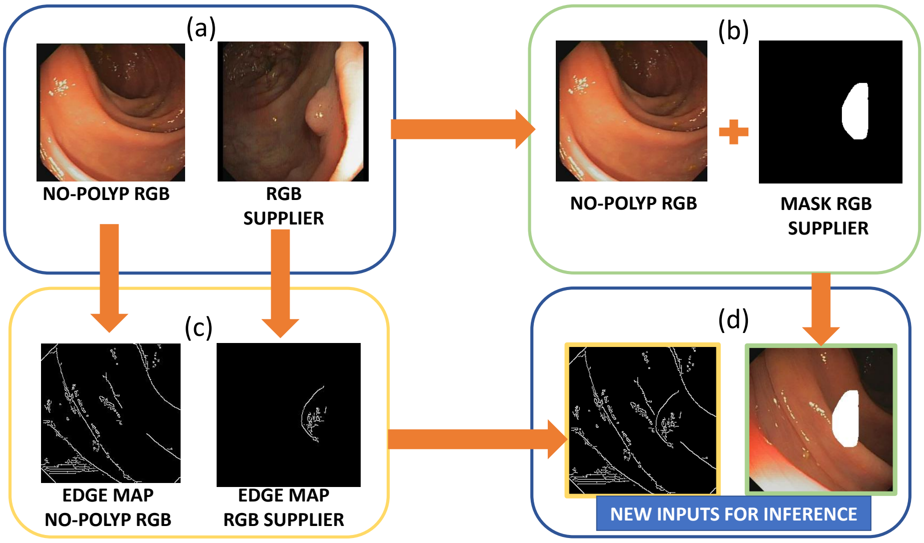

Figure 6.3: (a) We selected no-polyp frames from CVC-ClinicVideoDB and random rgb (RGB supplier) from Validation Set and Test Set of CVC-CLINIC. (b) We combined no-polyp RGB with the Mask of the selected RGB supplier. (c) We created the edge map of the no-polyp RGB and the edge map of the RGB supplier. We used Canny. (d) We combined the two edge maps from (c) to obtained the new edge map for inference (on the left). We also obtained from (b) the new RGB frame for inference (on the right).

where Rn and Pn are the precision and recall at the math th threshold.

In order to measure detection performance we employ a similar form of the usual AP, namely COCO's AP@[.5:.95] which depends on multiple IOUs (Intersection over union) and according to [88] it is a good trade-off to evaluate detection performance. For the sake of clarity, we remind that an IOU of 0.5 can be interpreted as an approximate identification of one of the detectable object, while an IOU of 1.0 is equivalent to a perfect localization of the such object. For a more strict evaluation concerning the likeness of the ground truth and detection bounding boxes, we use the AP@.50; this means that the IOU threshold to distinguish between true positive and false positive results is set to $t = 0.5$.

AR is an evaluation metric which measures the assertiveness of item detectors for a given class. Unlike AP, the confidences of the detection is not taken into account when calculating AR. Specifically it evaluates all the recall values obtained for IOU thresholds in the range $[0.5, 0.95]$ in steps of 0.1 and then it averages them (i.e., COCO AR@$[0.5, 0.95]$).

We set the maximum number of possible objects detectable using COCO metrics to evaluate as generic as possible and repeatable in other performances with other datasets.

6.2.3 Experiments and Results

In order to prove the validity of our proposal we compare the performance achieved with both, original dataset and the same dataset which has been augmented by using the proposed method. This highlights the improvement our inpainting data augmentation led. In summary, we have three kind of experiments: the first type consists in training the detector with CVCDb dataset without any data augmentation; in the second type of experiments we train the detector by using the same training set and by adding synthetic polyp images generated with the proposed method; in the third experiments type we further increase the training set by adding images from a different dataset, namely CVC. Finally, all the experiments are performed with three different detectors (i.e., Faster-RCNN, RetinaNet, EfficientDet) to prove the performance improvements does not depend on the model architecture. All the results are evaluated in term of AP@[0.5] and AR@[0.5:0.95]. An example of result is shown in Fig. 6.4.

Experiments without data augmentation

These baseline experiments were performed by splitting the CVCDb dataset into 70% for the training Set and 30% for the test set. Results for the three detectors

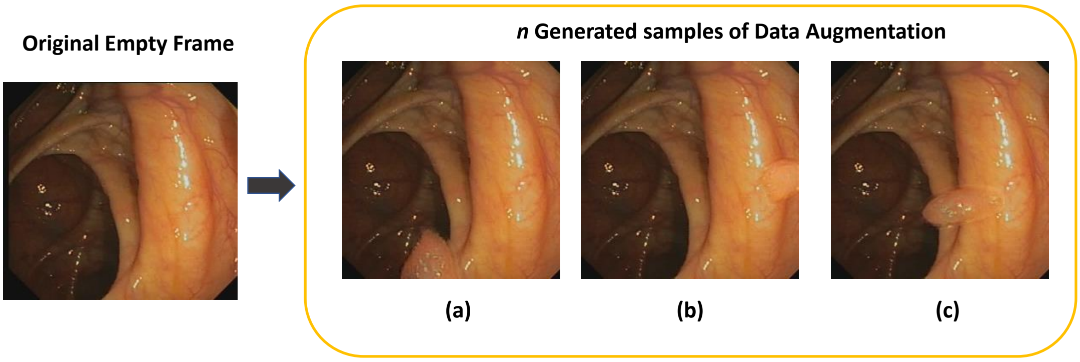

Figure 6.4: Example of results with our data augmentation contribution. The proposed methodology consists of generating n samples (right) from an empty frame (left). Three new frames (a), (b), (c) with polyps are noted in the example.

area reported in Table 6.1.

Experiments with proposed data augmentation

In this experiment we perform four tests with Faster-RCNN by adding 500, 1000, 4500 and 6000 generated polyps images. This allows us to show that higher is the number of synthetic images included, higher is the improvements in term of both, AP and AR. We observe and increase of 2.97% in AP and 0.7% in AR by augmenting training set with just 500 generated images. Increasing the numerosity of generated images shows a not surprising improvement until 9.98% in AP and 4.1% in AR for 6000 additional synthetic frames. This trend is clearly observable in in Table 6.2.

Experiments with proposed data augmentation and additional dataset

Here we increase the data by considering another dataset (Fig. 6.5). Specifically, we add the CVC data (i.e, 612 polyps images) to the training set we employed in the previous experiments, i.e., 70% of CVCDb plus the generated frames. Our aim is to increase training set variability, since the CVC images are different by the CVCDb ones in term of illumination, scale and sharpness of depicted polyps. Moreover, each CVC image is provided twice and in two forms: the original form, namely an unaltered frame; synthetic form, where we replaced the original polyp by using the proposed generator. Hence, we finally add to the previous training set, 612 original polyps images and 612 synthetic ones.

Table 6.3 shows the better performance we achieved with this setup with Faster-RCNN detector. We repeated the experiments with 1000, 4500 generated polyps images and add frames from CVC in the two aforementioned forms. The comparison with the baseline experiments (no augmentation) exhibits an improvement of 9.41% and 1.6% for AP and AR respectively. Performances further improve with 4500 generated images (+11.97% for AP and +3.3% for AR).

6.2.4 Tests with different detectors

To demonstrate that the proposed method does not depend on the detector, the previous three tests were performed but with only 1000 generated (in Table 6.4). With the following configurations an improvement in AP (+7.3%) and AR (+1.1%) is observed in the case of Faster R-CNN. While for RetinaNet an increase in AP (+4.7%) and AR (+1.9%) is observed. In the case of EfficientDet, the experiment with the addition of 1000 generated presents a (+6.7%) improvement for AP@.50, while a (+2.6%) improvement for AR. With the addition of samples from the other

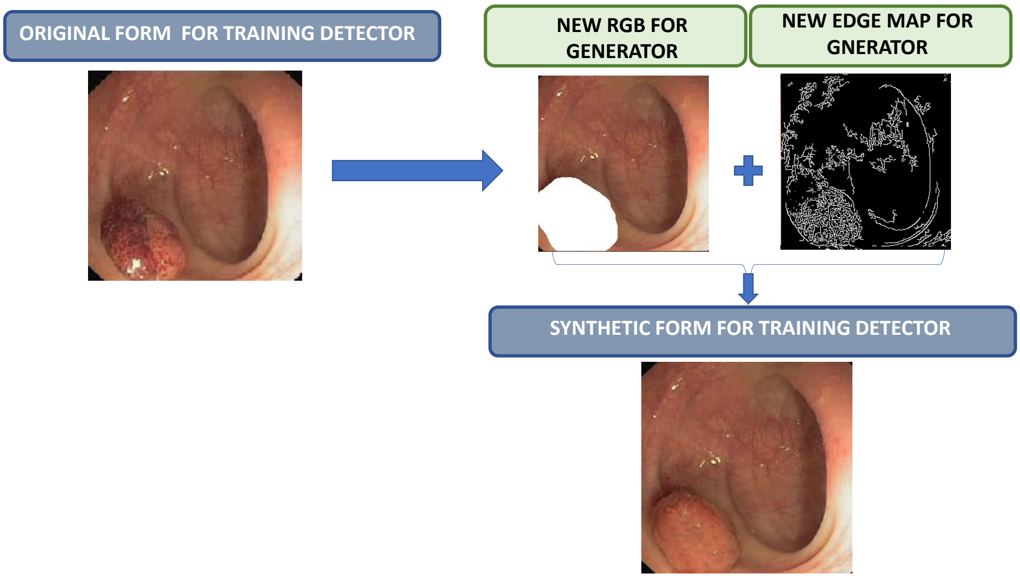
ORIGINAL FORM FOR TRAINING DETECTOR
NEW RGB FOR GENERATOR
NEW EDGE MAP FOR GNERATOR
SYNTHETIC FORM FOR TRAINING DETECTOR

Figure 6.5: Additional dataset for training detector. We increased the data by considering the CVC dataset, in original form and synthetic form.

Table 6.1: Experiment without data augmentation.

Model	**AP@.50**	**AR@[.50:.95]**
CVCDb	62.23	46.6

domain, there is also an increase of (+9.4%) for AP and (+1.6%) for AR. In the case of RetinaNet, while for EfficientDet there is an increase of +5.8 for AP and (+3.9) for AR. there is an improvement of (+8.1%) for AP, while a (+4%) for AR. The increase that is reported in the last two lines of the table show that regardless of the detector that is used, the methodology of data augmentation proposed in this paper, records an improvement in the task of polyp detection, then increasing the numerosity and at the same time the variability of the dataset, can help the training phase of a detector and then record better performance. This experiment was carried out with different detectors to show that whatever detector is used, works better in detection.

Table 6.2: Experiments with proposed data augmentation.

Model	**AP@.50**	**AR@[.50:.95]**
CVCDb + 500	65.20	47.3
CVCDb + 1000	69.50	47.7
CVCDb + 4500	68.94	50.2
CVCDb + 6000	72.22	50.7

Table 6.3: Experiments with proposed data augmentation and additional dataset.

Model	**AP@.50**	**AR@[.50:.95]**
CVCDb + CVC + 1000	71.27	47.7
CVCDb + CVC + CVC Synth + 1000	71.64	48.2
CVCDb + CVC + CVC Synth + 4500	74.20	49.9

Table 6.4: Evaluation comparison with different detectors

Model	**Faster R-CNN**		**RetinaNet**		**EfficientDet**	
-	**AP**	**AR**	**AP**	**AR**	**AP**	**AR**
CVCDb	62.2	46.6	62.5	38.1	56.1	46.2
CVCDb + 1000	69.5	47.7	67.2	40.0	62.8	48.8
CVCDb + CVC + CVC Synth + 1000	71.6	48.2	68.3	42.0	64.2	50.2

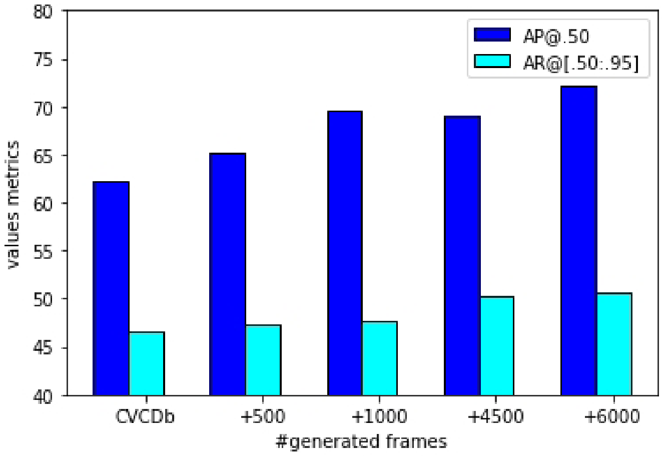

Figure 6.6: Barplot with trend of evaluation metrics.

6.3 Discussion

In this chapter, we presented a data augmentation methodology to be performed to generate a more numerous and variable dataset, useful for more robust training of any detector. This methodology exploits a known inpainting network, which has been adapted to our colonoscopy case. The main idea is to perform this step to improve the performance of polyp detection. The results obtained with Faster R-CNN show that a greater improvement is obtained as we increase the number of samples generated and integrate another data domain (Fig. 6.6). Experiments conducted with RetinaNet and EfficientDet also confirm this contribution. Therefore, this technique can be used with any other detector and having an even higher number of samples than the ones we have presented, a better result can be obtained.

Chapter 7

Conclusion and Future Works

In this work, we investigated the problem of automatic polyp detection in colonoscopies. We built a strong background on the domain and deep learning methods suitable to solve the problem addressed in this thesis. Through a domain analysis, it was easy to understand that, despite having a large dataset with high variability, there were several variables to consider in order to obtain a good result in polyp detection, such as: the presence of water bubbles in the frames, little difference between mucosa and polyp depending on the texture, various artifacts in the frames of colonoscopies. A first attempt to improve the performance in polyp detection was to specialize a neural network with clustering of polyp features. Results showed that such specialization with hierarchical fine-tuning can improve performance, but many false negatives remain to be attended to. This evidence suggested that we analyze these frames and introduce an attention mechanism within each video sequence. This proposed methodology resulted in better performance by lowering the number of false negatives. After this satisfactory result, we shifted our attention to a complete knowledge of the open-source datasets available from state-of-the-art. We noticed the low availability of large datasets with wide data variability. Therefore, we considered the creation of a framework capable of performing data augmentation through inpainting on colonoscopic data. Such a framework may be helpful to other researchers in this domain, who will be able to create larger and more realistic datasets in order to train any detector. This will decrease the over-fitting of the data when training any deep learning methodology.

Chapter 8

Appendix A

8.1 Other Publications

In the following, it is reported a work published during my Ph.D. but not directly related to this thesis.

International Conference:

- Pappalardo, G., Allegra, D., Stanco, F., & Battiato, S. (2019, September). A new framework for studying tubes rearrangement strategies in surveillance video synopsis. In 2019 IEEE international conference on image processing (ICIP) (pp. 664-668). IEEE.

Bibliography

[1] A. Sonnenberg and R. M. Genta. "Low prevalence of colon polyps in chronic inflammatory conditions of the colon". In: *Official journal of the American College of Gastroenterology— ACG* 110.7 (2015), pp. 1056–1061.

[2] D. K. Rex, D. A. Johnson, J. C. Anderson, P. S. Schoenfeld, C. A. Burke, and J. M. Inadomi. "American College of Gastroenterology guidelines for colorectal cancer screening 2008". In: *Official journal of the American College of Gastroenterology— ACG* 104.3 (2009), pp. 739–750.

[3] L. Villarosa. "Done Right, Colonoscopy Takes Time, Study Finds. Dec. 2006". In: *URL: http://www. nytimes. com/2006/12/19/health/19colo. html* ().

[4] M. Zorzi, L. Dal Maso, S. Francisci, C. Buzzoni, M. Rugge, and S. Guzzinati. "Trends of colorectal cancer incidence and mortality rates from 2003 to 2014 in Italy". In: *Tumori Journal* 105.5 (2019), pp. 417–426.

[5] 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)". In: *The American journal of gastroenterology* 98.2 (2003), pp. 471–477.

[6] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado. "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.

[7] J. Redmon and A. Farhadi. “Yolov3: An incremental improvement”. In: *arXiv preprint arXiv:1804.02767* (2018).

[8] S. Friederich. “Fine-tuning”. In: (2017).

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. “Generative adversarial nets”. In: *Advances in neural information processing systems* 27 (2014).

[10] M. Mirza and S. Osindero. “Conditional generative adversarial nets”. In: *arXiv preprint arXiv:1411.1784* (2014).

[11] J. Bernal, N. Tajkbaksh, F. J. Sánchez, B. J. Matuszewski, H. Chen, L. Yu, Q. Angermann, O. Romain, B. Rustad, I. Balasingham, et al. “Comparative validation of polyp detection methods in video colonoscopy: results from the MICCAI 2015 endoscopic vision challenge”. In: *IEEE Transactions on Medical Imaging* 36.6 (2017), pp. 1231–1249.

[12] Q. Angermann, A. Histace, and O. Romain. “Active learning for real time detection of polyps in videocolonoscopy”. In: *Procedia Computer Science* 90 (2016), pp. 182–187.

[13] K Geetha and C Rajan. “Automatic colorectal polyp detection in colonoscopy video frames”. In: *Asian Pacific journal of cancer prevention: APJCP* 17.11 (2016), p. 4869.

[14] N. Ahmed, T. Natarajan, and K. R. Rao. “Discrete cosine transform”. In: *IEEE transactions on Computers* 100.1 (1974), pp. 90–93.

[15] R. Zhang, Y. Zheng, C. C. Poon, D. Shen, and J. Y. Lau. “Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker”. In: *Pattern Recognition* 83 (2018), pp. 209–219.

[16] K. Pogorelov, O. Ostroukhova, M. Jeppsson, H. Espeland, C. Griwodz, T. de Lange, D. Johansen, M. Riegler, and P. Halvorsen. “Deep learning and hand-crafted feature based approaches for polyp detection in medical videos”. In: *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE. 2018, pp. 381–386.

[17] 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.

[18] 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.

[19] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. “Support vector machines”. In: *IEEE Intelligent Systems and their applications* 13.4 (1998), pp. 18–28.

[20] S. Y. Park, D. Sargent, I. Spofford, K. G. Vosburgh, A Yousif, et al. “A colon video analysis framework for polyp detection”. In: *IEEE Transactions on Biomedical Engineering* 59.5 (2012), pp. 1408–1418.

[21] 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.

[22] 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". In: *IEEE journal of biomedical and health informatics* 21.1 (2016), pp. 41–47.

[23] 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". In: *IEEE journal of biomedical and health informatics* 21.1 (2016), pp. 65–75.

[24] 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". In: *International journal of biomedical imaging* 2017 (2017).

[25] 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.

[26] O. Bardhi, D. Sierra-Sosa, B. Garcia-Zapirain, and A. Elmaghraby. "Automatic colon polyp detection using Convolutional encoder-decoder model". In: *2017 IEEE international symposium on signal processing and information technology (ISSPIT).* IEEE. 2017, pp. 445–448.

[27] 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". In: *Computerized Medical Imaging and Graphics* 43 (2015), pp. 99–111.

[28] Q. Angermann, J. Bernal, C. Sánchez-Montes, M. Hammami, G. Fernández-Esparrach, X. Dray, O. Romain, F. J. Sánchez, and A. Histace. "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.

[29] Y. Mori, S.-e. Kudo, M. Misawa, K. Takeda, T. Kudo, H. Itoh, M. Oda, and K. Mori. "Artificial intelligence for colorectal polyp detection and characterization". In: *Current Treatment Options in Gastroenterology* 18.2 (2020), pp. 200–211.

[30] C. Shorten and T. M. Khoshgoftaar. "A survey on image data augmentation for deep learning". In: *Journal of Big Data* 6.1 (2019), pp. 1–48.

[31] J. Kang and J. Gwak. "Ensemble of instance segmentation models for polyp segmentation in colonoscopy images". In: *IEEE Access* 7 (2019), pp. 26440–26447.

[32] Q. Nguyen and S.-W. Lee. "Colorectal segmentation using multiple encoder-decoder network in colonoscopy images". In: *2018 IEEE first international conference on artificial intelligence and knowledge engineering (AIKE).* IEEE. 2018, pp. 208–211.

[33] S. Mathew, S. Nadeem, S. Kumari, and A. Kaufman. "Augmenting Colonoscopy using Extended and Directional CycleGAN for Lossy Image Translation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2020, pp. 4696–4705.

[34] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *Proceedings of the IEEE international conference on computer vision.* 2017, pp. 2223–2232.

[35] P. Sasmal, M. Bhuyan, S. Sonowal, Y. Iwahori, and K. Kasugai. "Improved Endoscopic Polyp Classification using GAN Generated Synthetic Data Augmentation". In: *2020 IEEE Applied Signal Processing Conference (ASPCON).* IEEE. 2020, pp. 247–251.

[36] M. Arnold, A. Ghosh, S. Ameling, and G. Lacey. "Automatic segmentation and inpainting of specular highlights for endoscopic imaging". In: *EURASIP Journal on Image and Video Processing* 2010 (2010), pp. 1–12.

[37] J. Bernal, J. Sánchez, and F. Vilarino. "Impact of image preprocessing methods on polyp localization in colonoscopy frames". In: *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).* IEEE. 2013, pp. 7350–7354.

[38] S. M. Alsaleh, A. I. Aviles, P. Sobrevilla, A. Casals, and J. K. Hahn. "Adaptive segmentation and mask-specific Sobolev inpainting of specular highlights for endoscopic images". In: *2016 38th annual international conference of the ieee engineering in medicine and biology society (embc).* IEEE. 2016, pp. 1196–1199.

[39] P. Kazemi and I. Danaila. "Sobolev gradients and image interpolation". In: *SIAM Journal on Imaging Sciences* 5.2 (2012), pp. 601–624.

[40] M. Akbari, M. Mohrekesh, K. Najariani, N. Karimi, S. Samavi, and S. R. Soroushmehr. "Adaptive specular reflection detection and inpainting in colonoscopy

video frames". In: *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE. 2018, pp. 3134–3138.

[41] 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.

[42] 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.

[43] 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.

[44] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya. "A survey on addressing high-class imbalance in big data". In: *Journal of Big Data* 5.1 (2018), pp. 1–30.

[45] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[46] J. Prusa, T. M. Khoshgoftaar, D. J. Dittman, and A. Napolitano. "Using random undersampling to alleviate class imbalance on tweet sentiment data". In: *2015 IEEE international conference on information reuse and integration*. IEEE. 2015, pp. 197–202.

[47] S. Shekarforoush, R. Green, and R. Dyer. “Classifying commit messages: A case study in resampling techniques”. In: *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2017, pp. 1273–1280.

[48] Y. Zhu, C. Jia, F. Li, and J. Song. “Inspector: a lysine succinylation predictor based on edited nearest-neighbor undersampling and adaptive synthetic oversampling”. In: *Analytical biochemistry* 593 (2020), p. 113592.

[49] L. Bao, C. Juan, J. Li, and Y. Zhang. “Boosted near-miss under-sampling on SVM ensembles for concept detection in large-scale imbalanced datasets”. In: *Neurocomputing* 172 (2016), pp. 198–206.

[50] M. Kubat, S. Matwin, et al. “Addressing the curse of imbalanced training sets: one-sided selection”. In: *Icml*. Vol. 97. Citeseer. 1997, pp. 179–186.

[51] J. Laurikkala. “Improving identification of difficult small classes by balancing class distribution”. In: *Conference on Artificial Intelligence in Medicine in Europe*. Springer. 2001, pp. 63–66.

[52] 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.

[53] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. “The pascal visual object classes challenge: A retrospective”. In: *International journal of computer vision* 111.1 (2015), pp. 98–136.

[54] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. “Imagenet large scale visual recognition challenge”. In: *International journal of computer vision* 115.3 (2015), pp. 211–252.

[55] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. "The pascal visual object classes (voc) challenge". In: *International journal of computer vision* 88.2 (2010), pp. 303–338.

[56] Q.-C. Mao, H.-M. Sun, Y.-B. Liu, and R.-S. Jia. "Mini-YOLOv3: real-time object detector for embedded applications". In: *IEEE Access* 7 (2019), pp. 133529–133538.

[57] J. L. Elman. "Finding structure in time". In: *Cognitive science* 14.2 (1990), pp. 179–211.

[58] S. Hochreiter and J. Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[59] M.-T. Luong, H. Pham, and C. D. Manning. "Effective approaches to attention-based neural machine translation". In: *arXiv preprint arXiv:1508.04025* (2015).

[60] R. J. Williams. "Simple statistical gradient-following algorithms for connectionist reinforcement learning". In: *Machine learning* 8.3 (1992), pp. 229–256.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] Medium.com. *A simple overview of RNN, LSTM and Attention Mechanism*. `https://medium.com/swlh/a-simple-overview-of-rnn-lstm-and-attention-mechanism-9e844763d07b`. Online.

[69] K. Nazeri, E. Ng, T. Joseph, F. Z. Qureshi, and M. Ebrahimi. “Edgeconnect: Generative image inpainting with adversarial edge learning”. In: *arXiv preprint arXiv:1901.00212* (2019).

[70] K. He, X. Zhang, S. Ren, and J. Sun. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[71] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. “Image-to-image translation with conditional adversarial networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1125–1134.

[72] D. Ulyanov, A. Vedaldi, and V. Lempitsky. “Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6924–6932.

[73] S. Ren, K. He, R. Girshick, and J. Sun. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *Advances in neural information processing systems* 28 (2015), pp. 91–99.

[74] R. Girshick. “Fast R-CNN”. In: *International Conference on Computer Vision (ICCV)*. 2015.

[75] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. “Focal loss for dense object detection”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[76] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. “Feature pyramid networks for object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2117–2125.

[77] M. Tan, R. Pang, and Q. V. Le. “Efficientdet: Scalable and efficient object detection”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10781–10790.

[78] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. “Ssd: Single shot multibox detector”. In: *European conference on computer vision*. Springer. 2016, pp. 21–37.

[79] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[80] M. Tan and Q. Le. “Efficientnet: Rethinking model scaling for convolutional neural networks”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 6105–6114.

[81] A. I. Bandos, H. E. Rockette, T. Song, and D. Gur. “Area under the free-response ROC curve (FROC) and a related summary index”. In: *Biometrics* 65.1 (2009), pp. 247–256.

[82] Y. Shin, H. A. Qadir, and I. Balasingham. “Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance”. In: *IEEE Access* 6 (2018), pp. 56007–56017.

[83] J. Canny. “A computational approach to edge detection”. In: *IEEE Transactions on pattern analysis and machine intelligence* 6 (1986), pp. 679–698.

[84] R. Girshick. “Fast r-cnn”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448.

[85] Z. Liu, P. Luo, X. Wang, and X. Tang. “Deep learning face attributes in the wild”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 3730–3738.

[86] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. "Places: A 10 million image database for scene recognition". In: *IEEE transactions on pattern analysis and machine intelligence* 40.6 (2017), pp. 1452–1464.

[87] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. Efros. "What makes paris look like paris?" In: *ACM Transactions on Graphics* 31.4 (2012).

[88] R. Padilla, W. L. Passos, T. L. Dias, S. L. Netto, and E. A. da Silva. "A comparative analysis of object detection metrics with a companion open-source toolkit". In: *Electronics* 10.3 (2021), p. 279.