Multi-pathology Detection and Lesion Localization in WCE Videos by Using the Instance Segmentation Approach

Pedro M. Vieira[1], Nuno R. Freitas[1], Veríssimo B. Lima[1,2], Dalila Costa[3,4,5], Carla Rolanda[3,4,5], and Carlos S. Lima[1]

[1] CMEMS-UMinho Research Unit, Universidade do Minho, Guimarães, Portugal

[2] School of Engineering (ISEP), Polytechnic Institute of Porto (P.PORTO), Porto, Portugal

[3] Life and Health Sciences Research Institute, University of Minho, Campus Gualtar, 4710-057, Braga, Portugal

[4] ICVS/3Bs - PT Government Associate Laboratory, Braga/Guimarães, Portugal

[5] Department of Gastroenterology, Hospital de Braga, Braga, Portugal

Abstract

The majority of current systems for automatic diagnosis considers the detection of a unique and previously known pathology. Considering specifically the diagnosis of lesions in the small bowel using endoscopic capsule images, very few consider the possible existence of more than one pathology and when they do, they are mainly detection based systems therefore unable to localize the suspected lesions. Such systems do not fully satisfy the medical community, that in fact needs a system that detects any pathology and eventually more than one, when they coexist. In addition, besides the diagnostic capability of these systems, localizing the lesions in the image has been of great interest to the medical community, mainly for training medical personnel purposes. So, nowadays, the inclusion of the lesion location in automatic diagnostic systems is practically mandatory.

Multi-pathology detection can be seen as a multi-object detection task and as each frame can contain different instances of the same lesion, instance segmentation seems to be appropriate for the purpose. Consequently, we argue that a multi-pathology system benefits from using the instance segmentation approach, since classification and segmentation modules are both required complementing each other in lesion detection and localization. According to our best knowledge such a system does not yet exist for the detection of WCE pathologies.

This paper proposes a multi-pathology system that can be applied to WCE images, which uses the Mask Improved RCNN (MI-RCNN), a new mask subnet scheme which has shown to significantly improve mask predictions of the high performing state-of-the-art Mask-RCNN and PANet systems. A novel training strategy based on the second momentum is also proposed for the first time for training Mask-RCNN and PANet based systems. These approaches were tested using the public database KID, and the included pathologies were bleeding, angioectasias, polyps and inflammatory lesions. Experimental results show significant improvements for the proposed versions, reaching increases of almost 7% over the PANet model when the new proposed training approach was employed.

Keywords: Multi-pathology detection, Lesion localization, Instance segmentation, Mask-RCNN, Wireless Capsule Endoscopy, PANet.

1. Introduction

Wireless Capsule Endoscopy (WCE) is a diagnostic tool that has revolutionized the ability of visualization of the gastrointestinal tract. The classic endoscopy technique only reaches duodenum (the first section of the small bowel), and colonoscopy is limited to the last section of ileum (area of the small bowel closer to the large intestine). The middle section of the small bowel, which is usually more than 4 meters long, is almost impossible to directly visualize without using WCE (Basar et al., 2012). Since the final WCE video usually has more than 50,000 images, the development of

automatic diagnostic systems has been of interest for the researchers in the field. Nevertheless, most developed systems are capable of only detecting one lesion at a time, which is not what physicians look for in the clinical practice. Given the diversity of pathologies found in the gastro intestinal tract, multi-pathology detection is perhaps the last frontier for the massive use of automatic diagnosis systems integrated with wireless capsule endoscopy (WCE) exams (Thomson et al., 2001). In fact, the assumption of only one or a very small number of pathologies made by the most current diagnosis systems has been the biggest obstacle to its popularization in diagnostic procedures.

There are a variety of lesions that can appear in the small bowel, but only a small number of lesions appear frequently in WCE exams. From these, some can be pointed out, like: angioectasias, polyps, tumors, ulcers or bleeding (Thomson et al., 2001).

Convolutional Neural Networks (CNNs) have proliferated significantly and dominated the literature of modern automatic diagnosis systems due to their impressive performance. Even when using small datasets, which is frequently the case in medical applications, CNN-based automatic diagnosis systems can be applied. This is possible when using the Transfer Learning (TL) approach, that allows the training of most part of the system in large datasets (of different applications) and the use of the smaller dataset to perform refinements in the network, achieving high performances (Litjens et al., 2017; Yadav and Jadhav, 2019)

According to current literature Magnetic Resonance Imaging (MRI) is the most frequently data used in CNN-based systems, being the segmentation the most specific use (Choi and Jin, 2016; Dou et al., 2017; Havaei et al., 2017; Kamnitsas et al., 2017; Moeskops et al., 2016; Payan and Montana, 2015; Rasti et al., 2017). Gastrointestinal tract has been subject of some studies using CNNs for identification purposes, such as cancer detection (Li et al., 2018) or colorectal polyp detection (Liu et al., 2016; Misawa et al., 2018; Urban et al., 2018; Zhang et al., 2019), and even using the powerful instance segmentation approach based on the YOLO-net (Zheng et al., 2018), however in this case only for classification purposes. It has also been used for segmentation and classification in esophageal images (Wu et al., 2021) or for classification of abnormal images in large datasets (Guo and Yuan, 2020). CNNs have also been used for gastrointestinal automatic diagnosis based on WCE data for polyp detection (Baopu Li et al., 2009), ulcer and bleeding classification (Liaqat et al., 2018), ulcer detection (Alaskar et al., 2019) and ulcer and erosion detection (Fan et al., 2018). Angioectasias were also correctly detected using CNNs in different studies (Shvets et al., 2018; Tsuboi et al., 2020; Yusuf et al.) However, none of these systems offer a multi-pathology detection adequate for the clinical practice nor provide detection and segmentation results simultaneously. Systems with segmentation modules have been highly valued by the medical community, since the lesion tissue can be automatically assessed so the diagnostics process is speeded up. Additionally, segmentation modules allow the training of medical personnel in lesion detection.

The CNN-based structure mostly used for medical image segmentation in 2D images is perhaps the U-net (Ronneberger et al., 2015) given its potential. In this configuration, the input image is down sampled through a traditional CNN before being up sampled using transpose convolutions, until it reaches its original size. One of the most fundamental characteristics of U-net is that it skips connections that concatenate features from the down sampling to the up sampling paths, which is however based on

the ideas of ResNet. Several U-net variants have been used in the segmentation of WCE images for angioectasia segmentation (Shvets et al., 2018), which was the winning solution for MICCAI 2017 Endoscopic Vision SubChallenge. As U-net is a fully convolutional network it does not have any classification sub-net, therefore presents some restrictions for multi-pathology applications.

Solutions for multi-pathology applications are very rare, however some approaches have been proposed even for WCE applications. In (Iakovidis et al., 2018) a CNN is used for frame classification (normal or abnormal) by using weakly annotated images (different lesions only annotated as abnormal). Salient points obtained from deeper CNN layers are then processed by a Deep Saliency Detection algorithm in order to localize GI anomalies using an Iterative Cluster Unification algorithm. Although this system can deal with multi-pathology conditions, the type of pathology is never detected, therefore it cannot be considered as an automatic diagnosis system for multi-pathology applications. The system only provides the most likely lesion location in abnormal frames without considering its nature. For distinguishing the different types of lesions, a powerful segmentation module is missing as well as a sub-net for the detection of different pathologies. These are just the ingredients of instance segmentation systems such as Mask-RCNN (He et al., 2017), YOLO (Redmon et al., 2016) and PANet (Liu et al., 2018).

This paper proposes the Mask Improved R-CNN (MI-RCNN), an enhanced version of both mask subnets found in Mask-RCNN and PANet models, for multi-pathology detection and lesion localization in WCE videos. While the classification sub-net of the Mask-RCNN and PANet works well, some misalignments between the predicted and ground truth (GT) masks are found. Therefore, the mask sub-net needs to be improved. The most recent attempts to improve the mask sub-net are the Mask Scoring RCNN (MS-RCNN) (Huang et al., 2019) and the Boundary Mask-RCNN (BMask-RCNN) (Cheng et al., 2020). MS-RCNN adds a sub-net to the baseline Mask-RCNN to learn the quality of the predicted instance masks. The approach of BMask-RCNN (Cheng et al., 2020) is based on boundary information captured from the lower level of the pyramidal structure that enters to the new boundary-preserving mask sub-net, in which object boundary and mask are mutually learned via feature fusion blocks.

Instead of trying to add information that comes from different substructures of the Mask-RCNN to the mask sub-net (Huang et al., 2019), or merging mask sub-net information with information from the Region of Interest (RoI) alignment sub-net (Cheng et al., 2020), we propose making a more efficient use of the information that already exists in the mask subnet from the RoI alignment sub-net. This approach is followed in (Liu et al., 2018) where tiny fully connected layers which hold complementary properties to Fully Convolutional Network (FCN) (used in Mask-RCNN), can capture different views of each proposal. This increases the information diversity, hence producing masks of better quality. Several approaches show that propagating low-level information and combining information from different levels makes better use of the entire available information as observed in pyramidal based structures such as the Feature Pyramid Network (FPN), which is present in both Mask-RCNN and PANet. This approach, that was taken into account at the feature extraction level, was inexplicably ignored at the subnet mask module. The proposed MI-RCNN model improves the quality of predicted masks by propagating low-level information and combining information from different levels in the mask subnet proposed in PANet (Liu et al., 2018), which is already an improved version of the mask subnet proposed in

Mask-RCNN (He et al., 2017).

One of the most used methods for Neural Network training is the well-known Stochastic Gradient Descent (SGD) algorithm with momentum. The idea behind the momentum is to take advantage of the convergence dynamics on past iterations to more accurately predict the next one. Momentum codes directly the velocity of the weights variation along the training process. This reasoning can be extended by using higher order moments such as the second (acceleration) and third (variation of the acceleration) moments (Freitas et al., 2020). Second and third order moments can take advantage of previous information, which is not taken into account by the momentum. According to our best knowledge, higher order momentums have never been used in the training of neural networks. Convergence analysis and detailed explanation of the method can be found in (Freitas et al., 2020).

The developed system intends to simultaneously detect pathologies in WCE images, while delineating the abnormal tissue using an improved mask alignment and also a new method to accelerate convergence in neural networks training. Results reported in this paper constitute the use of higher order moments for the training of Mask-RCNN/PANet based models for the first time. Therefore, new training methodologies for state-of-the-art instance segmentation models are also proposed in the ambit of this paper. The output of the proposed system is the lesion localization (inside a bounding box) in each test image with indication of the most likely type of lesion. The system can detect different types of lesion or different pieces of the same type of lesion in the same frame. The proposed methodology was tested in the KID dataset (Iakovidis et al., 2018; Koulaouzidis et al., 2017).

2. Methodology

The methodology sectionproposed in this paper has three main sections: background, model structure and model training. The background section describes the different models that inspired the proposed MI-RCNN, while in. The model structure the approach to improve the mask sub-net of both the Mask-RCNN and PANet models is presented. Mask predictions can be improved by propagating and concatenating low and high layer information at the C4_fc layer (see Fig. 1) of the proposed PANet structure (Liu et al., 2018). Regarding model training, the use of the second momentum was proposed, which considers acceleration and velocity of the weight parameters instead of only the velocity considered by the regular (or first-order) momentum.

2.1. Background

When looking at MS-RCNN compared to Mask-RCNN, improvements can be seen in Average Precision (AP) of more than 1% in the COCO dataset. However, when deepening the results' analysis, we can see that the Mask-RCNN outperforms MS-RCNN for lower APs (Huang et al., 2019). This may be related with increased difficulties in the tracking of small objects since heuristically it seems to be more likely to have higher APs for large objects. Our results seem to confirm this suspicion, once that no AP improvements were obtained by using the MS-RCNN in the current application, perhaps because some of the lesions present in the small bowel are very small (e.g. angioectasias). Preliminary experiments show that PANet outperforms MS-RCNN for the KID dataset, therefore the focus was to improve PANet results by improving mask predictions which is the weak point of the original model. However, regarding to this approach two points must be taken into consideration:

1. Although in (Huang et al., 2019) is not referred a direct comparison between MS-RCNN and PANet using the demanding COCO dataset, when looking at both papers ((Liu et al., 2018) and (Huang et al., 2019)), it is possible to conclude that PANet shows improvements in the performance when applied to COCO dataset. This was the main reason to explore the PANet sub-mask module structure in this paper.
2. It is referred by MS-RCNN authors that there is room for mask prediction improvements in their method, which goes from 2.2% to 2.6%, depending on the used backbone (Huang et al., 2019). These results were found by changing the predicted mask for the GT in the training process. This approach seems promising however disregarded in the ambit of this paper since for the current implementation MS-RCNN underperforms PANet.

When looking at BMask-RCNN, the conclusion that can be made is that the method works well, however the data flow in the mask subnet is intricate and complex, and consequently hard to improve. Additionally, authors do not clearly separate and analyze improvements due to information propagation from lower pyramidal levels from improvements due to mask subnet changes. Better improvements are reported in the COCO dataset when compared with MS-RCNN (1.2 and 1.5 % respectively), however both methods present similar behaviors since AP improvements are higher for higher APs. It was also previously proven that for APs lower than 0.5 the improvement of BMask-RCNN over Mask-RCNN becomes negligible, becoming perhaps worsen for smaller APs (Cheng et al., 2020). These results clearly show poor boundary refinement in small objects.

2.2. Model Structure

The Mask R-CNN is perhaps the most popular CNN based structure for instance segmentation. Based on Fast/Faster R-CNN (Girshick, 2015; Ren et al., 2017), a FCN is used for mask prediction, along with box regression and classification. To achieve high performance, FPN is used to extract in-network feature hierarchy, where a top-down path with lateral connections is augmented to propagate semantically strong features. Practical applications show that this method sometimes can provide high classification scores associated with misaligned masks. This happens since the confidence of instance classification is used as a mask quality score in most instance segmentation frameworks. The mask quality, which is quantified as the Intersection over Union (IoU) between the instance mask and its ground truth, is usually not well correlated with classification score. As masks are predicted by a subnet, specifically conceived for this purpose, in principle mask improvements will be associated with the improvement of this network substructure.

Several improvements regarding mask predictions have been proposed, where the most prominent can be found in (Cheng et al., 2020; Huang et al., 2019; Liu et al., 2018). Despite the approaches proposed in (Cheng et al., 2020; Huang et al., 2019) make sense and proved to be effective in the very difficult COCO dataset, according with our findings both show not to be the best for medical image segmentation, especially if small lesions appear. Therefore, in this paper it was tried to improve the approach proposed in (Liu et al., 2018), in which the basic approach is to make better use of the available information from the RoI alignment module instead of merging information from different modules.

2.2.1. *Motivation*

Methods proposed in (Cheng et al., 2020; Huang et al., 2019; Liu et al., 2018) were developed to improve the mask predictions of the Mask-RCNN network. BMask-RCNN (Cheng et al., 2020) is the only that presupposes incomplete information at the mask subnet structure, since information propagation from low levels of the FPN is proposed. This information comes to the mask subnet from an alternate channel and is mixed by the information coming from the RoI alignment module by an intricate scheme along the FCN pipeline. This scheme, which is poorly explained, is therefore hard to understand and hard to improve. In this way, it wasn't an inspiration for us.

The MS-RCNN (Huang et al., 2019) is inspired by the idea of controlling directly the mask production, inserting a loss for the mask quality (which requires a new network branch named MaskIoU subnet allowing end-to-end training). The maskIoU subnet uses information from the most likely predicted mask along with information coming from the RoI alignment module, therefore only existing information in the mask subnet is used. In the PANet model (Liu et al., 2018), the proposed subnet mask fuses predictions from two views; the conventional FCN (which exists in Mask-RCNN) and small fully-connected layers, which possess complementary properties of FCN. These complementary properties are the core of the method since information diversity is increased, improving the quality of produced masks. Similarly to MS-RCNN, no information from other modules are propagated to the mask module, while assuming that all the required information for producing better masks exists in the mask subnet. However, this feature needs to be improved so it can achieve greater performances. This is exactly the approach followed in the MI-RCNN network proposed in this paper. We argue that enough information for producing better masks exists, however the mask subnet structure must be improved in order to make a better use of the existing information. Regarding FCN structures it is well known that forwarding low layer information and mixing this information with high level information makes a better use of the existing information. The best example of this is probably the FPN, which is based on the fact that low-level and high-level information complement each other and together provide more information than just high-level information alone. Several authors have also used fused feature maps for segmentation with finer details (Ghiasi and Fowlkes, 2016; Peng et al., 2017; Pinheiro et al., 2016), which is the method required to improve the masks produced by state-of-the-art instance segmentation systems.

2.2.2. *Proposed Approach*

In order to preserve most of the information along the CNN pipeline, it was proposed to join in the C4_fc layer information from all the preceding layers of the mask sub-net. The proposed structure is shown in Fig. 1 emphasizing the improvement of the Mask Head (mask sub-net) over the approach proposed in (Liu et al., 2018). In the first stage (backbone), FPN extracts features to generate RoIs via Region Proposal Network (RPN) for classification and bounding box regression, as it was proposed in (He et al., 2017). The backbone consists of a CNN which extracts features from raw images. This work followed the typical approach when using small datasets of fine-tuning ResNet-101-FPN, which was pre trained in the large COCO dataset. In this regard, the first four ResNet stages were frozen and only the weights of the last were trained with the remaining network. The mask sub-net uses each RoI features via RoIAlign, which preserves spatial information, essential for predicting segmentation. In fact, Faster R-

CNN uses RoIPool, which introduces misalignments between the RoI and the extracted features due to the consecutive quantizations, therefore losing some spatial information from the original images. RoiAlign aligns the extracted features with the input, preserving spatial information. (Yang et al., 2018).

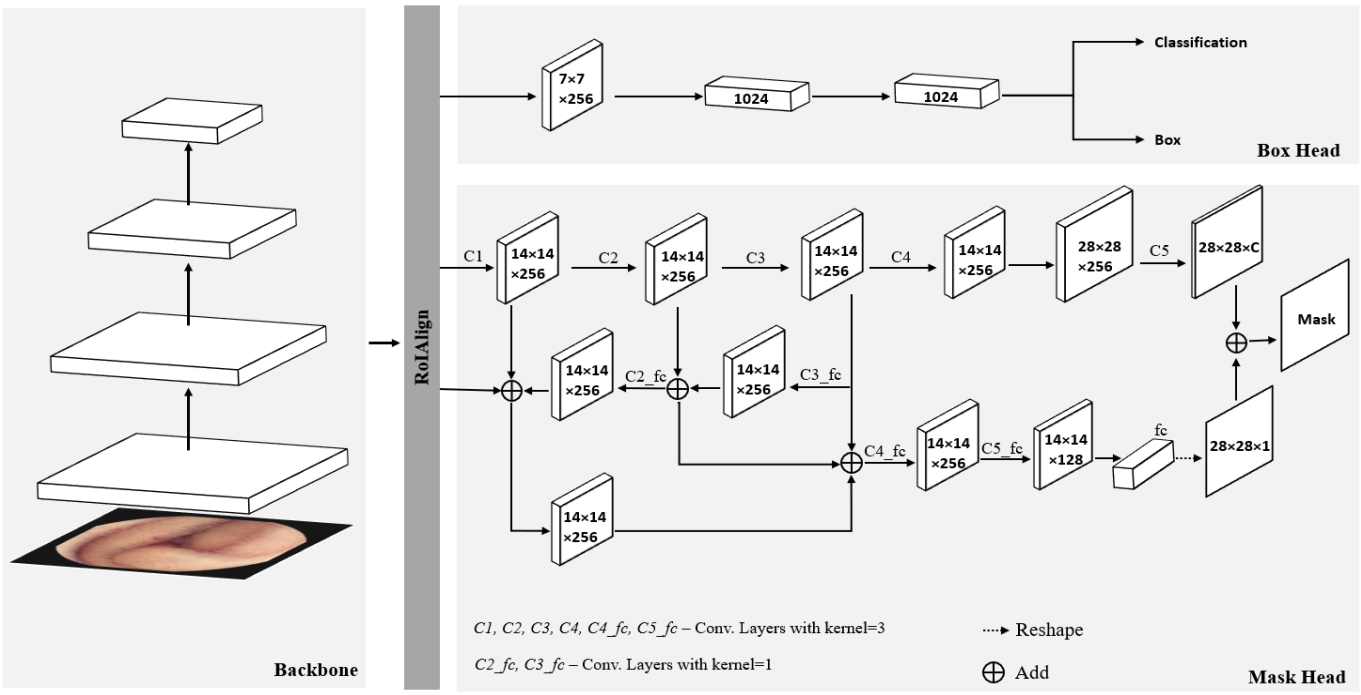

Fig. 1. Schematic diagram of the proposed MI-RCNN structure for the mask sub-net proposed in this paper.

2.3. *Model Training*

The standard gradient descent (SGD) update rule is given by:

$$w^{k+1} = w^k - \eta^k \nabla w^k f\left(w^k\right) \tag{1}$$

where *f(w)* is the function to be minimized, *k* stands for iteration number and η^k is the learning rate parameter. The minimization of *f*(*w*) can be accelerated by the SGD with momentum method. Momentum is given by:

$$z^{k+1} = \beta^k z^k + \eta^k \nabla w^k f\left(w^k\right)$$

$$w^{k+1} = w^k - z^{k+1} \tag{2}$$

where β^k is an iteration dependent parameter. Details on how the η^k and β^k parameters must be updated can be found in (Freitas et al., 2020) and references therein. The combination of the pair of equations (2) results in a new update rule given by the momentum:

$$w^{k+1} = w^k - \beta^k z^k - \eta^k \nabla w^k f\left(w^k\right) \tag{3}$$

By comparing equations (1) and (3) we can see that the momentum inserted the term $(-\beta^k z^k)$ in the update rule. By using the last equation of the pair of equations (2), equation (3) can be rewritten as:

$$w^{k+1} = w^k + \beta^k (w^k - w^{k-1}) - \eta^k \nabla w^k f\left(w^k\right) \tag{4}$$

Therefore, the term inserted by the momentum is the first difference of the network

weights weighted by the β^k parameter that must be adjusted.

Nesterov Accelerated Gradient (NAG) (Nesterov, 1983) computes equation (4) on the basis of the estimate of the next position of the parameters instead of on the current position. In addition to the momentum, NAG also significantly accelerates the algorithm convergence and the updating rule becomes:

$$w^{k+1} = w^k + \beta^k(w^k - w^{k-1}) - \eta^k \nabla w^k f\left(w^k - \beta^k z^k\right) \quad (5)$$

The acceleration of convergence through the momentum was based on the hypothesis that the successive aggregation of past gradient information is more effective than the latest negative gradient alone. In fact, the step taken at the previous iterate w^{k-1} was based on negative gradient information at that iteration, along with the search direction from the iteration prior to that one, w^{k-2}. By following this line of reasoning, we see that the previous step is a linear combination of all the gradient information found at all iterates so far, going back to the initial iterate w^0.

2.3.1. Higher Order Moments

Gradient information of past iterates is then given by the derivative of the weight parameters, which presupposes that at each iteration acceleration coefficients (second derivative of the weight parameters) can also encode convergence dynamics more extent in time and can help to improve convergence. A similar reasoning can be made regarding higher than second order moments. As the n^{th} order derivative is just the derivative of the $(n\text{-}1)^{th}$ derivative then the n^{th} moment is the momentum of the $(n\text{-}1)^{th}$ moment. Therefore the second momentum, which is the acceleration of the weight coefficients can be obtained by the momentum of the momentum given in equation (2) and is given in equation (6) already with the inclusion of the NAG which was named as NAG2:

$$\begin{gathered} s^{k+1} = \gamma^k s^k + \eta^k \nabla w^k f\left(w^k - \beta^k z^k - \gamma^k s^k\right) \\ z^{k+1} = \beta^k z^k + s^{k+1} \\ w^{k+1} = w^k - z^{k+1} \end{gathered} \quad (6)$$

The updating rule becomes:

$$w^{k+1} = w^k + \beta^k\left(w^k - w^{k-1}\right) + \gamma^k s^k - \eta^k \nabla w^k f\left(w^k - \beta^k z^k - \gamma^k s^k\right) \quad (7)$$

From the last two equations of the set of equations (6) we obtain:

$$\gamma^k s^k = \gamma^k\left[w^k - \left(1 + \beta^k\right)w^{k-1} + \beta^k w^{k-2}\right] \quad (8)$$

Equation (8) shows that the second momentum also reinforce indirectly the first momentum which can be seen rewriting equation (7) by inserting equation (8). Equation (7) becomes:

$$w^{k+1} = w^k + \left(\beta^k + \gamma^k\right)\left(w^k - w^{k-1}\right) - \beta^k\gamma^k\left(w^{k-1} - w^{k-2}\right) - \eta^k \nabla w^k f\left(w^k - \beta^k z^k - \gamma^k s^k\right) \quad (9)$$

Convergence analysis of the second momentum is discussed in (Freitas et al., 2020).

3. Experimental Results and Discussion

The effectiveness of the proposed approach was evaluated in the public database KID dataset 2 (Iakovidis et al., 2018; Koulaouzidis et al., 2017). This contains WCE images obtained from the whole GI tract using different exams, all taken with

MiroCam® (IntroMedic Co., Seoul, Korea) capsules. These images have a resolution of 360×360 pixels and all were manually annotated and scrutinized by an international scientific committee (Koulaouzidis et al., 2017). These include 303 images of vascular anomalies (small bowel angioectasias, lymphangiectasias, and blood in the lumen), 44 images of polypoid anomalies (lymphoid nodular hyperplasia, lymphoma, Peutz-Jeghers polyps) and 227 images of inflammatory anomalies (ulcers, aphthae, mucosal breaks with surrounding erythema, cobblestone mucosa, luminal stenoses and/or fibrotic strictures, and mucosal/villous oedema).

The vascular anomalies class was divided in two (angioectasias and bleeding), since these two lesions have too many differences in color, size and texture. Also, it is important to refer that for the physicians, the treatment and follow-up of these two lesions are extremely different. In this way, two physicians of Hospital of Braga reviewed the different images and classified them into angioectasia or bleeding separately, and in the case that a classification was different between them, a consensus was reached afterwards. So, the final dataset used in this work was composed of four different lesion classes: angioectasias, bleeding, polyps and inflammatory lesions. Examples of the different lesions present in the dataset can be seen in Fig. 2 and the final number of images by lesion are present in Table I.

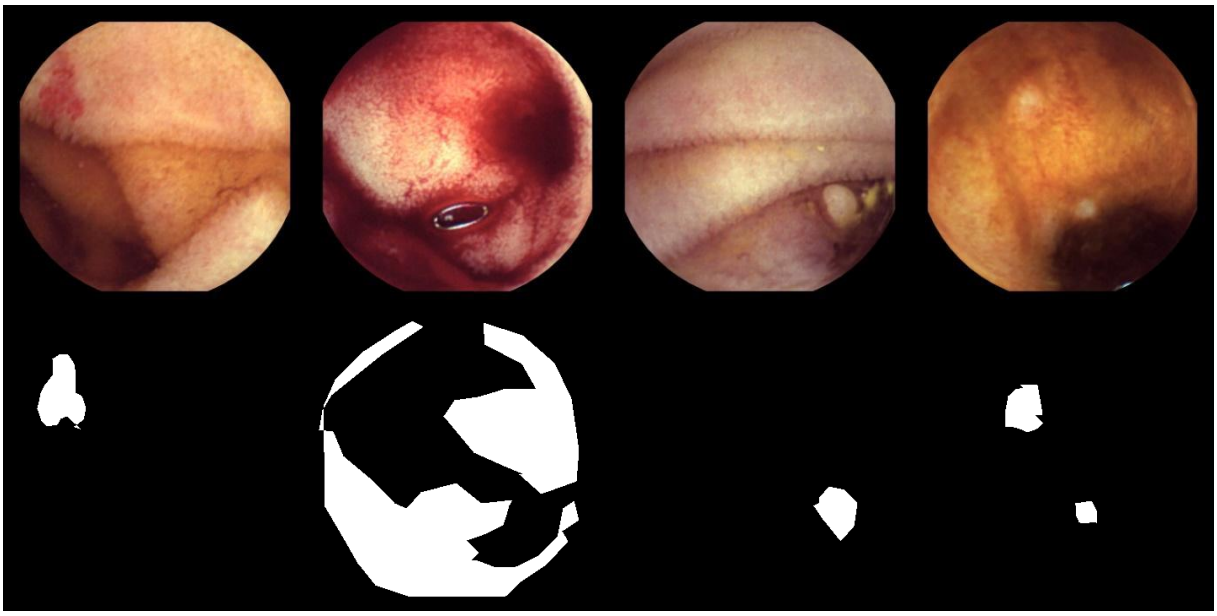

Fig. 2. Examples of lesions retrieved from KID Dataset 2. In the top the images and in the bottom the annotated masks. From the left to the right, an example of angioectasia, bleeding, polyp and inflammatory lesion.

Table I. Contents of the used dataset.

Lesion	Number of images
Normal	728*
Angioectasia	248
Bleeding	55
Polyps	44
Inflammatory	227

* Normal images were included only in the testing phase.

All frames were used in a proportion of 55% for training, 30% for testing and 15% for validation. All the sets were randomly selected from the entire dataset, since no subject related information is available in the KID dataset. As the background class is represented in all pathological frames, normal frames are not required for training purposes otherwise highly class imbalance will be obtained leading to a decreasing in performance.

Results were evaluated in terms of mean Average Precision (mAP), which is a metric often used in segmentation-based applications, and its two variants (mAP50 and mAP75), which returns the values of the precision considering different values of Intersection over Union (IoU). The F1-score metric was also included in this analysis, which uses both precision and recall for its computation. The considered baseline system is the Mask-RCNN (He et al., 2017). Table II shows the obtained results.

Table II. Comparison of the results obtained with the unseen test set among Mask R-CNN, PANet, the proposed approach with regular SGD and with the 2nd order moment optimizer in terms of mAP, mAP_{50}, mAP_{75} and F1-Score.

Method	Backbone	mAP	mAP_{50}	mAP_{75}	F1-score
Mask R-CNN (He et al., 2017)	ResNet-101 + FPN	33.10	56.51	34.71	49.23
PANet (Liu et al., 2018)	ResNet-101 + FPN	34.66	56.56	36.79	53.66
Proposed MI-RCNN	ResNet-101 + FPN	35.75	57.43	38.34	56.83
Proposed MI-RCNN + 2^{nd} moment	ResNet-101 + FPN	**40.35**	**59.42**	**43.01**	**60.07**

Regarding mask refinements, it is important to note that although outperforming the conventional Mask-RCNN in the COCO dataset, the very promising approaches proposed in (Cheng et al., 2020; Huang et al., 2019) both underperform the PANet for the KID dataset, therefore for clarity purposes were not shown in Table II. Results show that MI-RCNN with the regular SGD with momentum outperforms the baseline Mask R-CNN and PANet models, with improvements of almost 1% in mAP and mAP_{50}, 1.5% in mAP_{75} and more than 3% in F1-score. This shows the efficiency of propagating and mixing lower layer information with high level information in the mask subnet. By using the second momentum technique for training purposes, which is the second proposal of this paper and constitutes a novelty for a network of the size of the Mask-RCNN, an extra improvement was achieved. In this case, values of almost 5% in mAP, 2% in mAP_{50}, 5% in mAP_{75} and 3% in F1-score were obtained, which is consistent with the results presented in (Freitas et al., 2020). Globally, both proposals improve the results of the Mask-RCNN in the KID dataset, which is very significant. Although expecting a significant difference between mask scoring and classification results, the results in this case when looking at the masks shown a strong similarity between classification score and mask alignment, making the system more robust to changes.

When looking at the resultant masks (Fig. 3), it is visible that sometimes more than one lesion can be found (first and third example), but in all of these cases, the contour that better filled the annotated mask is the one with the higher probability returned by the model. It is important to note that in these cases, the region with the higher probability was considered to compute the metrics presented in Table II. It is also possible to understand that the examples with the smaller lesions are the ones where the resultant segmentation has the worst match when comparing to the mask, but still better than the state-of-the-art.

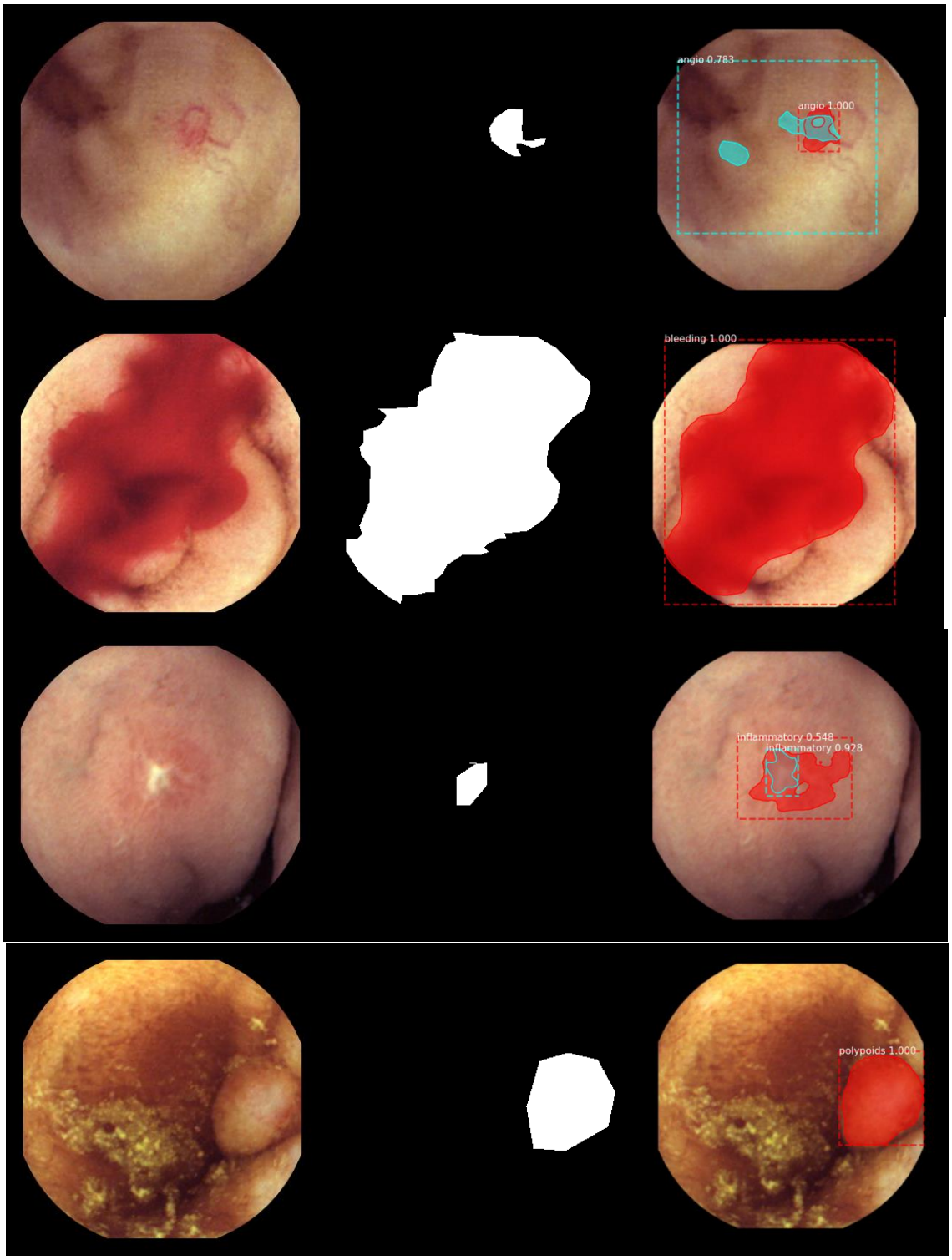

Fig. 3. Results of the segmentation using the proposed approach, in examples of the four different considered lesions. From the top to the bottom, an example of angioectasia, bleeding, polyp and inflammatory lesion. From the left to the right, the original image, the annotated mask and the predicted mask with the class probability.

The instance segmentation approach is a fundamental piece in automatic diagnosis for multi-pathology applications, however its use has been very limited which makes it difficult for performance comparisons with alternative systems as they do not exist. This is the case for multi-pathology detection and lesion segmentation by using WCE videos. In this regard the only possible comparison would be with (Iakovidis et al.,

2018), which is the most advanced known system that uses the KID dataset. However, as (Iakovidis et al., 2018) doesn't provide segmentation related metrics, only classification performances comparisons are possible. It is also important to notice that in (Iakovidis et al., 2018), although using the same dataset, the classification is done as normal/abnormal, and does not classify each lesion individually. Thus, it is not correct to directly compare our multi pathological system with the binary classification system in (Iakovidis et al., 2018). Table III shows the advantages of the instance segmentation approach. In fact, including both classification and segmentation modules in the same system, information sharing is promoted by different subsystems improving their joint use.

Table III. Comparison of the classification performances obtained with the unseen test set among Mask R-CNN, PANet, the proposed approach with regular SGD and with the 2nd order moment optimizer in terms of Recall, Precision, F1-score and Accuracy.

Method	Class	Recall	Precision	F1-score	Accuracy
Mask R-CNN (He et al., 2017)	Inflammatory	45.45	87.5	59.83	62.69
	Polypoids	71.43	90.91	80.00	
	Bleeding	50	**100**	66.67	
	Angioectasia	79.76	95.71	87.01	
	Weighted Average	62.69	92.49	73.76	
PANet (Liu et al., 2018)	Inflammatory	45.45	89.74	60.34	63.21
	Polypoids	78.57	84.62	81.48	
	Bleeding	61.11	84.62	70.97	
	Angioectasia	77.38	95.59	85.53	
	Weighted Average	63.21	91.44	73.83	
Proposed MI-RCNN	Inflammatory	51.95	90.91	66.12	66.32
	Polypoids	**78.57**	73.33	75.86	
	Bleeding	**66.67**	85.71	**75.00**	
	Angioectasia	77.38	94.20	84.97	
	Weighted Average	66.32	90.58	75.86	
Proposed MI-RCNN + 2nd moment	Inflammatory	**55.84**	**93.48**	**69.92**	**69.95**
	Polypoids	71.43	**100.00**	**83.33**	
	Bleeding	61.11	73.33	66.67	
	Angioectasia	**84.52**	**97.26**	**90.45**	
	Weighted Average	**69.95**	**93.72**	**79.52**	

Looking at Table III it is possible to conclude that the proposed MI-RCNN led to good performances in lesion detection, specially in the case of inflammatory, polypoids and

angioectasia lesions. It is noticeable a very promising result with angioectasia detection, which was not the most expected outcome due to the inclusion of other vascular-related lesions in the dataset (bleeding). On the other hand, the results of bleeding detection were substantially lower. By looking at the segmentation we can conclude that some of these lesions were mixed up with shadows and natural hollow spaces of the small bowel. It is also important to note that this was one of the less representative type of lesion in the dataset, which can damage the performance of the classifier. The polypoids detection always achieved the best results when MI-RCNN was applied, but a lower precision value was reached when applying the 2nd momentum, which was not expected. But looking at the overall performance, the 2nd momentum improved the results when looking at these lesions. Finally, when looking at inflammatory lesions, it is noticeable a significant difference between the precision and the recall values, but the proposed approach reached the best performance of the test. To compare these results with the performance of physicians, a 2012 study will be used (Zheng et al., 2012). In here, 17 WCE readers (8 with low experience and 9 with median or high experience) were used. In average, less than 50% of the lesions were found by this group of physicians; with a range that goes from 17% to 78%. Also, it is interesting to notice that there was no direct relationship between the reader experience in WCE analysis and the performance in this study (the physician with a performance of 17% was in fact the one with most experience in WCE). Regarding the different lesions present in the clips, angioectasias were the lesions most found (69%), followed by polyps (46%), ulcers (38%) and blood (17%). It is clear that in fact our system is able to classify a greater number of lesions in WCE abnormal frames, when compared to the average of physicians.

Regarding normal images, as explained previously, they were only included in the testing phase, since they are considered entirely background. Since the focus of this paper was to prove that different lesions in WCE images could be correctly classified , the results with normal images were not included into Table III. In fact, the True Positive (TP) rate when normal images are considered is 64.15% / 66.07% / 70.05% / 71.43%, when using Mask R-CNN, PANet, MI R-CNN and MI-RCNN with 2nd moment, respectively. It is clear that the proposed methodology improves the correct classification of normal images, outperforming sometimes physicians' performances hence able to be used in the clinical practice. If it is true that higher values are found in the literature, it is also true that they usually do not reflect current clinical practice with highly unbalanced classes and images hard to analyze.

Overall, when looking at the accuracy of the system, the MI-RCNN with the 2nd momentum reached the higher value, and it is a promising result, since no other previous work applied multi-pathology classification and segmentation in the KID dataset. Looking at (Iakovidis et al., 2018), the only work that used the whole KID dataset 2 for classification purposes, they have reached an accuracy of 77.5%. Although higher than our accuracy of 69.95%, results aren't comparable since (Iakovidis et al., 2018) used only a binary classification task of normal/abnormal. As a matter of fact, some pathological frames were incorrectly classified regarding its pathology, but this is not measured by (Iakovidis et al., 2018).

4. Conclusions

Current clinical practice of the gastrointestinal tract requires multi-pathology detection given the amount of different pathologies that can be found. Lesion localization

modules provided by modern automatic diagnosis systems have been highly appreciated by the medical community, since the specific region of the lesion is shown, improving the physician's confidence in the system. Therefore, a useful system must have a classification module and a segmentation module that can complement each other if they share components. Current instance segmentation systems have these characteristics, with Mask-RCNN being one of the most used systems of this type. One of the characteristics that can be improved in Mask-RCNN is the quality of the predicted mask. MS-RCNN and BMask-RCNN are two methods that improve the predicted mask quality in the COCO dataset, however both underperform PANet in our case. PANet adds a branch in the mask sub-net containing tiny fully connected layers that can capture different views of each proposal increasing information diversity hence producing masks of better quality.

This paper proposed MI-RCNN, an efficient method to improve the quality of the predicted mask that outperforms PANet. The method is based on the well-established principle that forwarding low-layer information and mixing this information with high-level information makes a better use of the existing information. In this regard, propagation of low layer information from all sub-net mask levels to the C4_fc layer is proposed. Also, the use of the 2nd momentum for training the network instead of a simple momentum was proposed, which is an innovative contribution from this work.

Experimental results show that the proposed methods significantly improve the evaluation metrics, with an increase of 3% in F1-score. By training the proposed model with the innovative method based on the second momentum an extra improvement of more than 3% was achieved over the PANet model. The classification results also followed the same behavior, with accuracies for the proposed MI-RCNN and MI-RCNN with 2nd momentum 3% and 7% higher than the PANet, respectively, for the majority of the analyzed lesions. Also, it is shown that the classification of normal images show improved performances when using the proposed methodology, with an increase of 4% and 6% in TP rate when comparing PANet with MI-RCNN and MI-RCNN with the 2nd momentum, respectively.

Although it is a first work in a scenario of multi-pathology detection in WCE images, the achieved results are quite promising. The good results not only in the classification task, but also on the segmentation task, could lead to the conclusion that in fact, these two modules should be always complementary to each other. As future work, we would like to improve the masks predictions to achieve better results, increase the number of pathologies by using other datasets and validate in a clinical setting the whole system.

Acknowledgments

This work was supported by FCT national funds, under the national support to R&D units grant, through the reference project UIDB/04436/2020 and UIDP/04436/2020 and through the PhD Grants with the references SFRH/BD/92143/2013 and SFRH/BD/139061/2018.

References

Alaskar, H., Hussain, A., Al-Aseem, N., Liatsis, P., Al-Jumeily, D., 2019. Application of Convolutional Neural Networks for Automated Ulcer Detection in Wireless Capsule Endoscopy Images. Sensors 19, 1265. https://doi.org/10.3390/s19061265

Baopu Li, Meng, M.Q.-H., Lisheng Xu, 2009. A comparative study of shape features for polyp detection in wireless capsule endoscopy images, in: 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, pp. 3731–3734. https://doi.org/10.1109/IEMBS.2009.5334875

Basar, M.R., Malek, F., Juni, K.M., Idris, M.S., Saleh, M.I.M., 2012. Ingestible Wireless Capsule Technology: A Review of Development and Future Indication. Int. J. Antennas Propag. 2012, 1–14. https://doi.org/10.1155/2012/807165

Cheng, T., Wang, X., Huang, L., Liu, W., 2020. Boundary-preserving Mask R-CNN, in: ECCV 2020. https://doi.org/10.1007/978-3-030-58568-6_39

Choi, H., Jin, K.H., 2016. Fast and robust segmentation of the striatum using deep convolutional neural networks. J. Neurosci. Methods 274, 146–153. https://doi.org/10.1016/j.jneumeth.2016.10.007

Dou, Q., Yu, L., Chen, H., Jin, Y., Yang, X., Qin, J., Heng, P.-A., 2017. 3D deeply supervised network for automated segmentation of volumetric medical images. Med. Image Anal. 41, 40–54. https://doi.org/10.1016/j.media.2017.05.001

Fan, S., Xu, L., Fan, Y., Wei, K., Li, L., 2018. Computer-aided detection of small intestinal ulcer and erosion in wireless capsule endoscopy images. Phys. Med. Biol. 63, 165001. https://doi.org/10.1088/1361-6560/aad51c

Freitas, N.R., P.M.Vieira, Vaz, A.I., Lima, C.S., 2020. Stochastic Optimization by using Higher-Order Moments. Manuscript Submitted for publication.

Ghiasi, G., Fowlkes, C.C., 2016. Laplacian Pyramid Reconstruction and Refinement for Semantic Segmentation. pp. 519–534. https://doi.org/10.1007/978-3-319-46487-9_32

Girshick, R., 2015. Fast R-CNN, in: 2015 IEEE International Conference on Computer Vision (ICCV). IEEE, pp. 1440–1448. https://doi.org/10.1109/ICCV.2015.169

Guo, X., Yuan, Y., 2020. Semi-supervised WCE image classification with adaptive aggregated attention. Med. Image Anal. 64, 101733. https://doi.org/10.1016/j.media.2020.101733

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., Larochelle, H., 2017. Brain tumor segmentation with Deep Neural Networks. Med. Image Anal. 35, 18–31. https://doi.org/10.1016/j.media.2016.05.004

He, K., Gkioxari, G., Dollar, P., Girshick, R., 2017. Mask R-CNN, in: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, pp. 2980–2988. https://doi.org/10.1109/ICCV.2017.322

Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X., 2019. Mask Scoring R-CNN, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 6402–6411. https://doi.org/10.1109/CVPR.2019.00657

Iakovidis, D.K., Georgakopoulos, S. V., Vasilakakis, M., Koulaouzidis, A., Plagianakos, V.P., 2018. Detecting and Locating Gastrointestinal Anomalies Using Deep Learning and Iterative Cluster Unification. IEEE Trans. Med. Imaging 37, 2196–2210. https://doi.org/10.1109/TMI.2018.2837002

Kamnitsas, K., Ledig, C., Newcombe, V.F.J., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. 36, 61–78. https://doi.org/10.1016/j.media.2016.10.004

Koulaouzidis, A., Iakovidis, D., Yung, D., Rondonotti, E., Kopylov, U., Plevris, J., Toth, E., Eliakim, A., Wurm Johansson, G., Marlicz, W., Mavrogenis, G., Nemeth, A., Thorlacius, H., Tontini, G., 2017. KID Project: an internet-based digital video atlas of capsule endoscopy for research purposes. Endosc. Int. Open 05, E477–E483. https://doi.org/10.1055/s-0043-105488

Li, Y., Li, X., Xie, X., Shen, L., 2018. Deep learning based gastric cancer identification, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, pp. 182–185. https://doi.org/10.1109/ISBI.2018.8363550

Liaqat, A., Khan, M.A., Shah, J.H., Sharif, M., Yasmin, M., Fernandes, S.L., 2018. Automated Ulcer and Bleeding Classification from WCE Images using Multiple Features Fusion and

Selection. J. Mech. Med. Biol. 18, 1850038. https://doi.org/10.1142/S0219519418500380

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88. https://doi.org/10.1016/j.media.2017.07.005

Liu, S., Qi, L., Qin, H., Shi, J., Jia, J., 2018. Path Aggregation Network for Instance Segmentation, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, pp. 8759–8768. https://doi.org/10.1109/CVPR.2018.00913

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C., 2016. SSD: Single Shot MultiBox Detector. pp. 21–37. https://doi.org/10.1007/978-3-319-46448-0_2

Misawa, M., Kudo, S., Mori, Y., Cho, T., Kataoka, S., Yamauchi, A., Ogawa, Y., Maeda, Y., Takeda, K., Ichimasa, K., Nakamura, H., Yagawa, Y., Toyoshima, N., Ogata, N., Kudo, T., Hisayuki, T., Hayashi, T., Wakamura, K., Baba, T., Ishida, F., Itoh, H., Roth, H., Oda, M., Mori, K., 2018. Artificial Intelligence-Assisted Polyp Detection for Colonoscopy: Initial Experience. Gastroenterology 154, 2027-2029.e3. https://doi.org/10.1053/j.gastro.2018.04.003

Moeskops, P., Viergever, M.A., Mendrik, A.M., de Vries, L.S., Benders, M.J.N.L., Isgum, I., 2016. Automatic Segmentation of MR Brain Images With a Convolutional Neural Network. IEEE Trans. Med. Imaging 35, 1252–1261. https://doi.org/10.1109/TMI.2016.2548501

Nesterov, Y., 1983. A Method for Solving a Convex Programming Problem with Convergence Rate O(1/K^2). Sov. Math. Dokl. 27, 372–376.

Payan, A., Montana, G., 2015. Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks.

Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J., 2017. Large Kernel Matters — Improve Semantic Segmentation by Global Convolutional Network, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1743–1751. https://doi.org/10.1109/CVPR.2017.189

Pinheiro, P.O., Lin, T.-Y., Collobert, R., Dollár, P., 2016. Learning to Refine Object Segments. pp. 75–91. https://doi.org/10.1007/978-3-319-46448-0_5

Rasti, R., Teshnehlab, M., Phung, S.L., 2017. Breast cancer diagnosis in DCE-MRI using mixture ensemble of convolutional neural networks. Pattern Recognit. 72, 381–390. https://doi.org/10.1016/j.patcog.2017.08.004

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You Only Look Once: Unified, Real-Time Object Detection, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 779–788. https://doi.org/10.1109/CVPR.2016.91

Ren, S., He, K., Girshick, R., Sun, J., 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans. Pattern Anal. Mach. Intell. 39, 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation.

Shvets, A., Iglovikov, V., Rakhlin, A., Kalinin, A.A., 2018. Angiodysplasia Detection and Localization Using Deep Convolutional Neural Networks. Cold Spring Harb. Lab. https://doi.org/10.1101/306159

Thomson, A.B.R., Keelan, M., Thiesen, A., Clandinin, M.T., Ropeleski, M., Wild, G.E., 2001. Small bowel review: normal physiology part 1. Dig. Dis. Sci. 46, 2567–2587.

Tsuboi, A., Oka, S., Aoyama, K., Saito, H., Aoki, T., Yamada, A., Matsuda, T., Fujishiro, M., Ishihara, S., Nakahori, M., Koike, K., Tanaka, S., Tada, T., 2020. Artificial intelligence using a convolutional neural network for automatic detection of small-bowel angioectasia in capsule endoscopy images. Dig. Endosc. 32, 382–390. https://doi.org/10.1111/den.13507

Urban, G., Tripathi, P., Alkayali, T., Mittal, M., Jalali, F., Karnes, W., Baldi, P., 2018. Deep Learning Localizes and Identifies Polyps in Real Time With 96% Accuracy in Screening Colonoscopy. Gastroenterology 155, 1069-1078.e8. https://doi.org/10.1053/j.gastro.2018.06.037

Wu, Z., Ge, R., Wen, M., Liu, G., Chen, Y., Zhang, P., He, X., Hua, J., Luo, L., Li, S., 2021. ELNet:Automatic classification and segmentation for esophageal lesions using convolutional neural network. Med. Image Anal. 67, 101838. https://doi.org/10.1016/j.media.2020.101838

Yadav, S.S., Jadhav, S.M., 2019. Deep convolutional neural network based medical image classification for disease diagnosis. J. Big Data 6, 113. https://doi.org/10.1186/s40537-019-0276-2

Yang, X., Sun, H., Fu, K., Yang, J., Sun, X., Yan, M., Guo, Z., 2018. Automatic Ship Detection in Remote Sensing Images from Google Earth of Complex Scenes Based on Multiscale Rotation Dense Feature Pyramid Networks. Remote Sens. 10, 132. https://doi.org/10.3390/rs10010132

Yusuf, S., Hawken, S., Ounpuu, S., Dans, T., Avezum, A., Lanas, F., McQueen, M., Budaj, A., Pais, P., Varigos, J., Lisheng, L.,. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. Lancet 364, 937–52. https://doi.org/10.1016/S0140-6736(04)17018-9

Zhang, X., Chen, F., Yu, T., An, J., Huang, Z., Liu, J., Hu, W., Wang, L., Duan, H., Si, J., 2019. Real-time gastric polyp detection using convolutional neural networks. PLoS One 14, e0214133. https://doi.org/10.1371/journal.pone.0214133

Zheng, Y., Hawkins, L., Wolff, J., Goloubeva, O., Goldberg, E., 2012. Detection of lesions during capsule endoscopy: physician performance is disappointing. Am. J. Gastroenterol. 107, 554–60. https://doi.org/10.1038/ajg.2011.461

Zheng, Y., Zhang, R., Yu, R., Jiang, Y., Mak, T.W.C., Wong, S.H., Lau, J.Y.W., Poon, C.C.Y., 2018. Localisation of Colorectal Polyps by Convolutional Neural Network Features Learnt from White Light and Narrow Band Endoscopic Images of Multiple Databases, in: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, pp. 4142–4145. https://doi.org/10.1109/EMBC.2018.8513337