# Detecting and Locating Gastrointestinal Anomalies Using Deep Learning and Iterative Cluster Unification

Dimitris K. Iakovidis*, *Senior Member IEEE*, Spiros V. Georgakopoulos, Michael Vasilakakis, Anastasios Koulaouzidis, and Vassilis P. Plagianakos, *Member IEEE*

***Abstract*—This paper proposes a novel methodology for automatic detection and localization of gastrointestinal (GI) anomalies in endoscopic video frame sequences. Training is performed with weakly annotated images, using only image-level, semantic labels instead of detailed, pixel-level annotations. This makes it a cost-effective approach for the analysis of large videoendoscopy repositories. Other advantages of the proposed methodology include its capability to suggest possible locations of GI anomalies within the video frames, and its generality, in the sense that abnormal frame detection is based on automatically derived image features. It is implemented in three phases: a) It classifies the video frames into abnormal or normal using a Weakly Supervised Convolutional Neural Network (WCNN) architecture; b) detects salient points from deeper WCNN layers, using a Deep Saliency Detection (DSD) algorithm; and c) localizes GI anomalies using an Iterative Cluster Unification (ICU) algorithm. ICU is based on a Pointwise Cross-Feature-Map (PCFM) descriptor extracted locally from the detected salient points using information derived from the WCNN. Results from extensive experimentation using publicly available collections of gastrointestinal endoscopy video frames, are presented. The datasets used include a variety of GI anomalies. Both the anomaly detection and the localization performance achieved, in terms of the Area Under receiver operating Characteristic (AUC), were >80%. The highest AUC for anomaly detection was obtained on conventional gastroscopy images, reaching 96%, and the highest AUC for anomaly localization was obtained on wireless capsule endoscopy images, reaching 88%.**

***Index Terms*—Endoscopy, gastrointestinal tract, computer-aided detection and diagnosis, machine learning.**

## I. Introduction

Gastrointestinal Endoscopy (GIE) is a fundamental modality for the investigation of the gastrointestinal (GI) tract and the detection of luminal pathology. The most common GIE procedures are gastroscopy and colonoscopy. Another GIE procedure, which has become the prime choice for the examination of the small bowel, is wireless capsule endoscopy (WCE) [1]. WCE is performed with a swallowable, untethered capsule equipped with a camera that captures color images during its journey along the GI tract. The amount of images produced during any GIE procedure is significantly large and with a diverse content, making the detection of GI anomalies a challenging task for image-based Medical Decision Support Systems (MDSS).

MDSS for GIE appeared primarily to cover clinical needs related to the detection and localization of lesions suspicious for malignancy or of bleeding sources, and to provide a second opinion on the assessment of lesions that require a more thorough examination [1–3]. This way, the application of such systems can contribute in speeding up the flexible endoscopy procedures, which are uncomfortable for a lot of patients, and can also enable less experienced personnel to perform it, including physicians' extenders or specialty nurses. Therefore, a consequent increase in clinical productivity, and an overall cost reduction for healthcare systems is possible [1]. The immense clinical need for such systems is more apparent in WCE. During a WCE video review, WCE readers usually reach their human limits by trying to maintain their concentration undistracted in order to examine approximately 50,000-120,000 images within an average of 60-90min [1]. This could explain the low diagnostic accuracy of WCE [4]. Relevant commercially available solutions at present include visualization enhancement algorithms for the discrimination of anomalies, such as the Flexible Spectral Colour Enhancement (FICE), and blood detection algorithms, such as the Suspected Blood Indicator (SBI) [1]. Other commercial, recently released software solutions incorporate methods for the detection and handling of uninformative images, e.g., identical images or images full of bubbles and debris, to speed up the review times in WCE [5].

The majority of current MDSS for GIE are based on supervised machine learning algorithms aiming to detect/diagnose possibly abnormal conditions in the medical images. They are usually trained with annotated images, in which the locations of anomalies associated with the abnormal conditions, is indicated. Typically, the training images are annotated by experts at pixel-level, i.e., the experts indicate which pixels correspond to anomalies.

In this paper we investigate weakly-supervised learning for automated video analysis in GIE. This type of machine learning has been investigated as an alternative to cope with the resource-demanding issue of detailed image annotation [6]. It involves annotation of the training images only at image-level, using a semantic tag indicating whether the

This paper was submitted for review on August, 1, 2017, and received in revised form on May 4, 2018. This work was supported in part by the project "**Klearchos Koulaouzidis**", Grant No. 5151, and Grant No. 5024 of the Special Account of Research Grants of the University of Thessaly, Greece.

D.K. Iakovidis, S.V. Georgakopoulos, V.P. Plagianakos and M. Vasilakakis, are with the Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, Greece (e-mail: {diakovidis, spirosgeorg, vasilaka, vpp}@uth.gr). A. Koulaouzidis is with the Endoscopy Unit of The Royal Infirmary of Edinburgh, Edinburgh, UK (e-mail: akoulaouzidis@hotmail.com).

image contains anomalies or not; thus, omitting the details that can be specified by pixel-level annotation.

We propose a novel methodology based on a deep, Convolutional Neural Network (CNN) architecture that includes saliency detection, and an Iterative Cluster Unification (ICU) algorithm. Unlike any previous weakly supervised GIE video analysis approaches, the proposed methodology provides both automatic image feature extraction and anomaly localization capabilities. It receives a whole video frame sequence as input and it outputs suggestions about the existence of GI anomalies along with their possible locations within the video frames. Contributions of this paper beyond the state-of-the-art include:

- A Weakly-supervised CNN (WCNN) –based methodology for both anomaly detection and localization in the context of GIE, using training samples annotated solely with image-level labels;
- A novel Deep Saliency Detection (DSD) algorithm, enabling the detection of salient points relevant to GI anomalies in endoscopic video frames;
- A novel ICU algorithm enabling the localization of the anomalies within the video frames, based on Pointwise Cross-Feature-Map (PCFM) WCNN features. These are extracted from the salient points detected by DSD;
- Application of the proposed methodology in the GIE domain, using publicly available datasets that include a diverse set of anomalies and normal video frames from various parts of the GI tract.

The rest of this paper consists of five sections. Section II provides an overview of the related state-of-the-art methods. The proposed methodology is presented in Section III. Section IV describes the datasets used in this study, and Section V presents the results obtained from the experimental evaluation of the proposed methodology on these datasets. A discussion and a summary of conclusions are provided in the last section.

## II. Related Work

The first MDSS for automated detection of GI anomalies in GIE video sequences appeared in the early 2000's [3]. Since then, a variety of such systems has been proposed, aiming to reduce the number of the lesions missed during GIE [7]. These mainly include supervised approaches addressing the detection of only a single or a few kinds of GI anomalies [2], including polyps [8–14], both ulcers and polyps [15], esophageal cancer [16], celiac disease [17], inflammatory lesions [18], [19], and bleeding [20–22]. Some recent approaches are more general in the sense that they address the detection of various kinds of anomalies. These include methods based on a Deep Sparse Support Vector Machine (DSSVM) and superpixel segmentation [23], [24], a method based on color saliency and Support Vector Machines (SVMs) [25], [26], and methods based on CNNs [27], [28].

CNNs are contemporary extensions of the well-known Multi-layer Feed-forward Neural Networks (MFNNs) characterized by a deep structure that enables feature extraction from raw input images through layers of adaptable filtering components [29]. This makes them independent from any hand-crafted feature extraction method "tailored" to specific diagnostic tasks [30]. They have been utilized in a variety of medical imaging domains both as conventional supervised classifiers, trained using image patches [8], [11–13], [27], [28], [31–34] and as weakly-supervised classifiers trained using weakly-annotated images [14], [18], [35–37]. Considering that image patches are sampled from known locations within the images, patch-based methods enable both the detection and the localization of possible anomalies; however, they require training with images annotated at pixel-level. In one of the most recent patch-based CNN approaches addressing the detection of various kinds of GI anomalies [27], input images were represented in CIE-Lab color space, and the CNN had a relatively low number of filters.

A preliminary study utilizing a CNN in a weakly-supervised framework (WCNN) was performed by our research group [18], aiming at the detection of inflammatory lesions. Recently, weakly supervised CNN-based approaches have been proposed in the context of GIE. These include a CNN that apart from the RGB images it receives their Hessian and Laplacian transformations as input [37]. A more complex cascaded CNN scheme was proposed for the recognition of the different organs of the GI tract and normal intestinal content [36]. A CNN architecture as in [18], but using an SVM in place of the second fully connected layer, has been proposed for the detection of blood in WCE [20]. In [14], accurate detection of polyps in white-light and narrow-band imaging endoscopy, was reported using a pre-trained CNN only as a feature extractor. The pre-training was performed with non-medical images from the ImageNet dataset. A standard SVM was used for the classification of the CNN feature vectors.

Recently, Bag of visual Words (BoW) has also been identified as an effective weakly-supervised machine learning strategy to cope with the demand for annotated training GIE images [19], [22], [38–40]. Features considered for the generation of the BoW vocabulary in these studies include color histograms extracted from various color spaces for bleeding detection [22]; CIE-Lab features [19] for inflammatory lesion detection; a combination of Scale-Invariant Feature Transform (SIFT) with complete Local Binary Pattern (CLBP) histograms for polyp detection [40]; and, a combination of various colour and LBP histograms for the detection of gastric and oesophageal cancer, gastritis, and oesophagitis [38]. In the latter approach the weak labels are automatically mined from diagnostic texts.

A drawback of most weakly-supervised approaches, over the patch-based ones, is that they do not provide information about the location of anomalies within an image. Only a few approaches have been proposed to this direction. State-of-the-art generic weakly supervised CNN-based methodologies with localization capabilities, have been proposed mainly in the context of classification and segmentation of real-world objects [41], [42]. The methodology proposed in [41] requires a pre-training stage, using images annotated at pixel-level, whereas the methodology proposed in [42] uses weakly-labeled images or sub-images as bounding boxes of the objects of interest. The latter is based on the DeepLab-CRF model [43], which

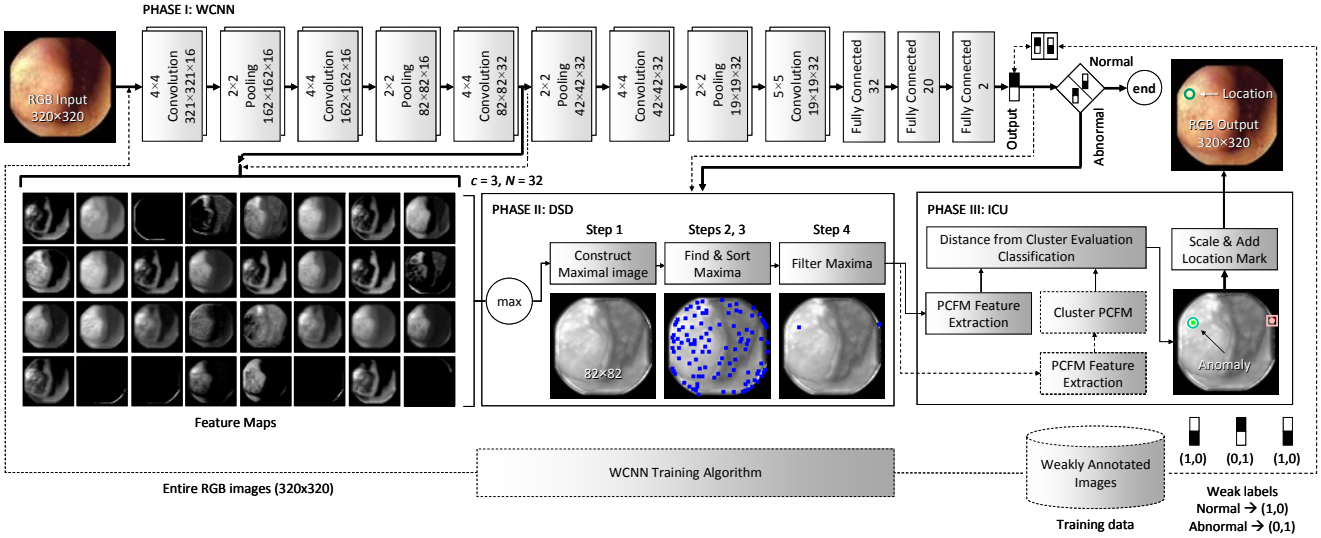


Figure 1. Proposed methodology. An endoscopic image is semantically characterized as abnormal or normal by the WCNN (Phase I). Abnormal images are further analyzed by DSD salient point detection algorithm (Phase II). The salient points are classified by ICU algorithm to identify and localize possible anomalies (Phase III). The dashed lines are used to indicate the workflow of the training process.

combines a CNN with a fully connected Conditional Random Field (CRF) aiming to the segmentation of the object of interest. In that study, it was shown that the use of weak annotations solely at the image-level is insufficient to train a high-quality segmentation model, and that the segmentation results become sufficient only when bounding boxes are used. The results improved with the use of pixel-level annotations from a subset of training images in a semi-supervised context.

In the context of GIE image analysis, anomaly localization has been based mainly on unsupervised image segmentation approaches, applicable on images with identified anomalies. As in the case of current anomaly detection methods most of them address the segmentation of only specific kinds of anomalies, such as polypoid lesions [3], [9], [44], and bleeding regions [21]. Recently, we have investigated the application of a localized region-based active contour model for the unsupervised segmentation of various kinds of lesions [45], and we highlighted its utility for measurement of lesion sizes. Lesion localization, as considered in the current study, aims to attract the attention of the video reviewer at specific points within an image, where anomalies are possibly located. The specification of a few points instead of the segmentation of whole image regions provides more targeted cues about the location of the anomalies, while it usually involves fewer computations; therefore, it is preferable in terms of time-efficiency for application on GIE video frame sequences.

The proposed approach exploits a WCNN architecture to detect and describe salient points within GIE images. In contrast to the current CNN-based image descriptors, which are mainly global [14], [46], [47], PCFM pixel-level descriptors are extracted from each salient point. A novel ICU algorithm utilizes these descriptors to discriminate pixels that correspond to suspicious image regions without any detailed, pixel-level annotation. Unlike state-of-the-art approaches its application is not limited to specific GI anomalies, and it is investigated across different GIE modalities, including WCE and gastroscopy.

## III. Proposed Methodology

The proposed methodology aims at the detection and localization of anomalies in images from GIE video. It is implemented in three phases, as illustrated in Fig. 1:

- Phase I: A deep WCNN architecture classifies the GIE images as abnormal (having GI anomalies) or normal;
- Phase II: DSD is applied on the abnormal images to detect salient points in the input images using information extracted from the feature maps of a deeper WCNN convolutional layer;
- Phase III: ICU is applied to identify a subset of salient points that possibly belong to GI anomalies. The coordinates of these points are then transformed (linearly scaled up) to match the spatial resolution of the input endoscopic image, on which they are superimposed to indicate the possible locations of the anomalies.

The application of this methodology presumes that that both WCNN and ICU are trained. The training workflow is indicated with dashed lines in Fig.1. WCNN is trained independently, using weakly annotated images (the square after the WCNN represents the loss calculated during training). DSD is applied on both abnormal and normal training images to detect salient points. Then ICU extracts PCFM features and uses them to form clusters, labeled as abnormal or normal. This process requires that training images are only weakly annotated. The formed clusters concentrate the information required for identification of points corresponding to anomalies within the unknown test images. The implementation details of these phases are provided in the following subsections, respectively.

### A. *Deep Classification of GIE Images*

A WCNN architecture is used for weakly-supervised learning of the GIE images. The images of the training set are weakly-annotated, i.e., image-level instead of pixel-level labels are used. The labels indicate whether an input image includes a GI anomaly or not, and they are represented by

respective binary vectors (0,1) and (1,0), at the two-neuron output layer of the WCNN architecture (Fig. 1).

The input images are of 24 bit RGB color with a size of 320×320 pixels (their size is limited by the memory available in the GPU); thus, larger images are downscaled to match these dimensions. The information loss due to downscaling is a compromise made to maintain the complexity of the architecture sufficiently low for average hardware requirements, without significantly affecting the appearance of the anomalies. The input image dimensions were kept fixed for generality purposes. Considering that the WCNN receives an entire image as input, any changes in the dimensions of the input images can have an impact on its overall architecture (e.g., the number of convolutional layers). The CNN is composed of five convolutional layers. Each of the first four convolutional layers is followed by a max-pooling layer, and the fifth convolutional layer is followed by three fully-connected layers (including the output layer). The pooling layers facilitate downscaling of the feature maps of the respective convolutional layers. The first four convolutional layers are composed of 4×4 kernel filters with stride 1 and padding 2, and the fifth convolutional layer has 5×5 kernel filters with stride 1 and padding 1. The pooling filters are 2×2 with stride 2 and no padding. The number of convolutional layers has been limited to five, because we found that it provides the best tradeoff between the depth of the network and the overall classification performance for the particular input image dimensions, i.e., the use of more than five convolutional layers results in a marginal improvement of the classification performance (of the order of $10^{-3}$), thus, considering our limited computational resources as well, we decided not to further increase the computational complexity by adding more layers. Each of the first two convolutional layers consists of 16 feature maps (empirically determined) followed by 16 max-pooling filters, while the next convolutional and pooling layers have 32 feature maps and max-pooling filters, respectively. The number of feature maps increases with the reduction of their spatial resolution, as suggested in [48]. The ReLU activation function is used in each convolutional layer. The first two fully-connected layers consist of 32 and 20 sigmoid (tanh) neurons, and the output layer has 2 softmax neurons respectively. The fully-connected layers facilitate the dimensionality reduction of the feature maps of the last convolutional layer for classification. The number of neurons was empirically determined. The selection of the output neurons was driven by the need to have the endoscopic images classified into two classes.

### B. Salient Point Detection

The feature maps of a convolutional layer *c* of the WCNN are used for the generation of an intermediate image from which a set of salient points is detected. This image is generated as a projection of the maximum values of all the feature maps $F_j^c$, $j$=1,2,…,$N$ of that layer, thus it is referred to as *maximal image* $M^c$. It has the same size as the feature maps, and each pixel value in location $(k,l)$ is estimated as the maximum of the respective pixel values of the feature maps $F_j^c$,

***Algorithm 1*** *Deep Saliency Detection (DSD)*

1. Construct a *maximal image* $M^c$ from a deeper WCNN convolutional layer *c* using Eq.(3);
2. Find the local maxima in $M^c$ using a maximum filter and add them in a list *L*;
3. Sort *L* in a descending order of intensities;
4. For each element *l* of *L* do:
   - Visit each element *k* of $M^c$ in the 8-connected neighborhood of *l* in $M^c$;
   - Initialize list $\Im$ by adding *l*;
   - If *valueOf(k)*$\in$(*valueOf*(*l*)-*t*, *valueOf(l)*] then *k'*=*k*, mark *k'* as "*candidate*", and recursively do:
     - If *valueOf(k')* = *valueOf(l)* then add *k'* as a "*valid maximum*" in $\Im$;
     - Visit and mark as "*candidate*" each 8-connected neighbor *k* of "*candidate*" *k'* until all *k* have *valueOf(k)* < *valueOf(l)-t*;
     - If *k* exists in *L* then remove *k* from *L*;
   - Produce a salient point by calculating the geometric center of all elements in $\Im$.

$$M^c(k,l) = \max\{F_j^c(k,l) \mid j = 1,2,..., N\}, \qquad (3)$$

where *j* is the feature map index of the convolutional layer *c*, and *N* is the number of the feature maps of that layer.

The selection of the convolutional layer *c* is driven by the capacity of its feature maps to highlight localized features of the anomalies. The deeper feature maps of a CNN tend to highlight such features [49]. However, after several pooling layers, the spatial resolution of very deep feature maps can become significantly smaller than that of the input images. The correspondences that can be established between such feature maps and the input images, by scaling, will consequently become very approximate, and the uncertainty of the anomaly localization task will increase. Therefore, a middle layer is expected to be more appropriate for the construction of the *maximal image*. Based on these considerations (experimentally validated in Section V.D) the third (*c*=3) WCNN convolutional layer is chosen.

By feeding the WCNN with an abnormal image, the points of the *maximal image* that have higher values are likely to represent GI anomalies, and vice versa. In that sense, the maxima of the maximal image correspond to salient points in the input image. The detection of the salient points is implemented using DSD (Algorithm 1). This algorithm begins in Step 1 by constructing the maximal image $M^c$. In Step 2 it detects the local maxima of $M^c$. The respective points are denoted as *l*. It stores these points in a list *L*, and sorts them in Step 3 upon their intensities, which are denoted as *valueOf*(*l*). In Step 4, it determines the maxima of $M^c$ that stand out from their surroundings by a value *t*, where *t* represents the tolerance of the algorithm to greylevel variations by controlling the extent of seed filling around maxima [50], [51]. This is done by visiting all the neighboring points of *l* in *L* and by marking as "*candidate*" those with an intensity value in the interval between *valueOf*(*l*)-*t* and *valueOf*(*l*). This process extends to the neighbors of the "*candidate*" points recursively

until no other neighboring points with values in this interval can be found. If an element of $L$ appears within the visited neighbors, it is removed from $L$ as a weaker maximum, and it is not further processed. From the points marked as "*candidate*", those that have an intensity value equal to the maximum $l$ of the current iteration are marked as "*valid maxima*" and they are added in the list $\Im$. The salient points are produced by calculating the geometric center of the "*valid maxima"*, collected in the list $\Im$, per iteration of Step 4.

An example is illustrated in Fig. 1 (Phase II). It can be noticed that initially, several maxima are detected and sorted in $L$ (Steps 2, 3). After Step 4 the maxima are drastically reduced to only two points. This can be explained by the relatively high tolerance value used ($t$=120). Generally, as the value of $t$ increases, the interval between *valueOf*($l$)-$t$ and *valueOf*($l$) becomes wider; it is, therefore, expected to have fewer salient points since more initial maxima are considered as weaker, and the calculation of the geometric center is performed with maxima of wider image regions.

### C. Anomaly Detection by Iterative Cluster Unification

The last phase of the proposed methodology aims to determine which salient points of the abnormal images discovered in the first phase (Section III.A), belong to GI anomalies. The proposed methodology assumes that abnormal images contain both abnormal and normal regions, whereas normal images contain only normal regions. Each salient point, detected using the DSD algorithm on these images, is represented by a feature vector composed of the values of each of the feature-maps derived from convolutional layer $c$ of the WCNN at this point. The dimensionality of the feature vector is the equal to the number of feature maps of convolutional layer $c$ (for $c$=3 the dimensionality is 32). The derived pointwise cross-feature-map (PCFM) features are used as input to the ICU classifier.

The ICU algorithm (Algorithm 2) is based on clustering to classify the salient points detected by the DSD algorithm in a weakly supervised way. It involves a training and a testing phase. During training it receives both abnormal and normal training images, and clusters their salient points upon their vector representations (which are unlabelled because the images are only weakly annotated). It considers that in the abnormal images some salient points may fall into normal regions as well. Iteratively, ICU unifies the clusters of the salient points of the abnormal images that are more similar to the clusters of the normal images. In the testing phase ICU receives an image classified as abnormal by WCNN (Phase I), and the detected salient points (Phase II) are classified into possibly abnormal or normal upon the $K$-nearest neighbor ($K$-NN) clusters in the unified cluster space.

For simplicity, the clustering algorithm used in this study is the well-known $k$-means algorithm [52]. Preliminary investigation using other clustering algorithms, including fuzzy c-means [52] and our recent random direction divisive clustering algorithm [53], did not lead to any significant classification performance improvement. To cope with the fact that the result of this algorithm depends significantly on its initialization, the clustering algorithm is performed for $T$ iterations with different initializations. Thus, a richer and more representative clustered vector space is generated by selecting $T > 1$. For the estimation of the distances between the clusters the Euclidean distance metric between the centroids of the clusters was used.

***Algorithm 2*** *Iterative Cluster Unification (ICU)*

*Training phase*

1. Let $I_n$ and $I_a$ be the training sets of normal and abnormal images respectively;
2. Let $Z_n$ and $Z_a$ be the sets of salient points extracted using DSD algorithm from $I_n$ and $I_a$;
3. For $i$ = 1 to $T$ do:
   - For each normal image in $I_n$ do:
     - Extract PCFM representations of $Z_n$;
   - Cluster the PCFM representations of $Z_n$ into $Q$ clusters $N_q$, $q = 1,2,\ldots,Q$;
   - For each abnormal image in $I_a$ do:
     - Extract PCFM representations of $Z_a$;
   - Cluster the PCFM representations of $Z_a$ into $R$ clusters $A_r$, $r = 1,2,\ldots,R$;
4. Set $Q = Q \cdot T$, $R = R \cdot T$;
5. For each abnormal cluster $A_r$, $r = 1,2,\ldots R$ do:
   - Calculate all distances $d_{rq}\left(A_r, N_q\right)$, $q = 1,2,\ldots Q$ between $A_r$ and $N_q$;
   - Sort distances $d_{rq}$ in ascending order;
   - Calculate the normalized distance $d_{rq12} = d_{rq1} / d_{rq2}$, where $d_{rq1}$ and $d_{rq2}$ represent the distances of $A_r$ to its closest neighboring clusters $N_{q1}$ and $N_{q2}$;
6. Estimate the mean normalized distance $d_{q12}$ from all $d_{rq12}, r = 1,2,\ldots R$ calculated in step 5;
7. For each abnormal cluster $A_r$, $r = 1,2,\ldots R$ do:
   - If $d_{rq12} < d_{q12}$ then unify $A_r$ with normal clusters: $Q$=$Q$+1; $N_Q = A_r$; $A_r = \varnothing$;

*Test phase*

1. Let $I_i^a$ be a new input image, characterized as abnormal by the WCNN classifier;
2. For each salient point $s$ in $I_i^a$ do:
   - Extract a PCFM representation of point $s$;
   - Calculate the distances of $s$ from all clusters in $A_r \cup N_q$;
   - Classify $s$ as normal or abnormal based on its $K$ nearest neighbors by majority voting;

## IV. Datasets

In order to enable reproducibility of the experiments and comparisons with current and future studies, two publicly available image datasets were used for the evaluation of the proposed WCNN architecture. These datasets have been acquired with different endoscopic imaging modalities. They have been selected primarily for their diversity, as they include different kinds of anomalies and normal images.

### A. MICCAI Gastroscopy Challenge Dataset

The first dataset considered in this study is composed of images obtained from gastroscopies. It was released for the

purposes of a challenge that took place in MICCAI 2015[1] [24]. The task in that challenge was to correctly classify the gastroscopic images and to detect abnormalities. In this paper, the same dataset is used for the detection of abnormal images using only semantically annotated training images.

The gastroscopy challenge dataset was derived from a total of 10,000 images, obtained from 544 healthy volunteers and from 519 volunteers with various lesions, such as gastritis, cancer, bleeding and gastric ulcer. The original image resolution was 768×576 pixels. The images were anonymized by cropping the image regions containing sensitive patient information. The size of the derived images is 489×409 pixels [24].

For the purposes of the MICCAI challenge, a subset of images was selected and separated, into two approximately balanced sets of training and a set of testing images. The training set consists of 205 normal and 260 abnormal images, and the test set consists of 104 normal and 129 abnormal images. We keep this separation so as to be able to compare our results. In what follows this will be referred to as Dataset 1.

### B. KID Dataset

Aiming to contribute to essential progress in the field of MDSS for WCE we have recently released KID[2] [45], a publicly available database of annotated WCE images and videos (including pixel-level annotations), which can be used as a reference for both training and evaluation of such systems [2], [45]. The second image dataset used in this study, is composed of images from KID Dataset 2. It contains WCE images obtained from the whole GI tract using a MiroCam capsule endoscope with a resolution of 360×360 pixels. These include 303 images of vascular anomalies (small bowel angiectasias, lymphangiectasias, and blood in the lumen), 44 images of polypoid anomalies (lymphoid nodular hyperplasia, lymphoma, Peutz-Jeghers polyps), 227 images of inflammatory anomalies (ulcers, aphthae, mucosal breaks with surrounding erythema, cobblestone mucosa, luminal stenoses and/or fibrotic strictures, and mucosal/villous oedema), and 1,778 normal images obtained from the esophagus, the stomach, the small bowel and the colon. In the rest of this paper, this 2,352 image dataset will be referred to as Dataset 2.

Following the paradigm of Dataset 1, and in order to be able to express the results using simple evaluation metrics, such as the classification accuracy (which depends on the distribution of the classes in the dataset), a balanced subset of the KID dataset was constructed. This was performed with random sub-sampling of the normal images to obtain an approximately equal number of normal and abnormal images. In the sequel, the dataset was divided into a training set with 429 normal and 423 abnormal images, and a test set with 172 normal and 172 abnormal images. In what follows, this dataset will be referred to as balanced Dataset 2 (Dataset 2B).

## V. Experiments and Results

Seven sets of experiments were conducted to evaluate the proposed methodology, using different parameters along its phases on different datasets. The first set of experiments investigated the image classification performance of the WCNN using various learning algorithms (Phase I). The best performing learning algorithm was used in the comparison of the WCNN with conventional and other weakly supervised approaches in the second and the third set of experiments respectively. The fourth set of experiments investigated the performance of the WCNN-based salient point detection approach (Phase II). The fifth set evaluated the performance of the proposed methodology overall, as it infers the localization of the GI anomalies in its output (Phase III). A sixth set of experiments evaluated the proposed methodology on different anomaly detection and localization tasks. Time-efficiency was investigated in the last set of experiments. The experimental procedures and the results obtained are presented in the following subsections.

TABLE I
Classification Accuracy of WCNN Using Different Learning Algorithms on Datasets 1 and 2B (D1-D2B)

| Algorithm | Mean D1 | St.D. D1 | Best D1 | Mean D2B | St.D. D2B | Best D2B |
|---|---|---|---|---|---|---|
| SGD | **0.836** | **0.048** | **0.909** | **0.874** | **0.007** | **0.892** |
| MSGD | 0.801 | 0.027 | 0.867 | 0.869 | 0.012 | 0.898 |
| Adam | 0.820 | 0.030 | 0.884 | 0.863 | 0.017 | 0.895 |

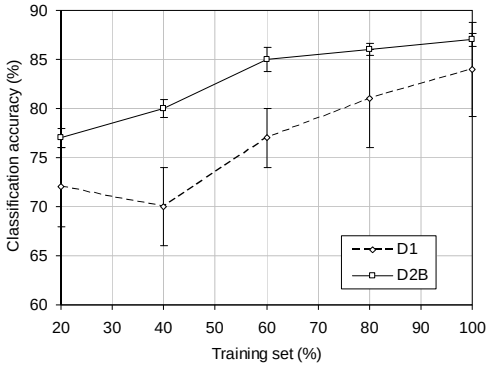


Figure 2. Classification performance of the proposed WCNN for various subsets of the training sets obtained from datasets 1 and 2B (D1-D2B).

### A. Investigation of WCNN Learning Algorithms

The generalization performance of a neural network can significantly vary upon the choice of the learning algorithm. This motivated a preliminary set of experiments which investigated the performance of the proposed WCNN architecture using three different learning algorithms, namely the original Stochastic Gradient Descent (SGD), the SGD with momentum (MSGD) [54] and Adam [55]. The latter is a state-of-the-art algorithm for first-order gradient-based optimization of stochastic objective functions, which uses adaptive estimates of lower-order moments.

We performed 100 executions for each learning algorithm on each of the available datasets. For experimentation efficiency, the classification accuracy was used as a performance measure because it is simple, intuitive, and sufficiently reliable, considering that the class distribution of the datasets is approximately balanced. The learning rate was

[1] Dataset 1: http://endovissub-abnormal.grand-challenge.org/
[2] Dataset 2: http://is-innovation.eu/kid/

the same and equal to 0.001 for all the learning algorithms. In all the cases the maximum number of training epochs was 500; however, in most cases approximately 100 epochs were sufficient for the algorithm to converge. An epoch is considered completed with the full pass of the training set, which is divided in mini-batches. Each batch consisted of 50 images. The momentum constant was set to 0.9 and the Adam learning algorithm was executed using the default parameters suggested in [55]. The results obtained are summarized in Table I, using the mean, standard deviation, and best scores obtained from the 100 executions of each learning algorithm on Datasets 1 and 2B. The SGD algorithm without momentum term provided the highest performance in both datasets. The lower performance of Adam could be justified by the fact that it may not always converge to optimal solutions [56].

Three statistical tests were performed to validate the significance of the classification performance differences presented in Table I. These include a Shapiro–Wilk normality test [57], a non-parametric Friedman test, and a two-sided Wilcoxon rank sum test [58]. In all tests the null hypothesis (i.e., that the samples are independent and derived by identical continuous distributions with equal medians) was rejected, justifying that there exist differences between the methods at the 5% significance level ($p$-values < 0.05).

The robustness of the results of the SGD algorithm was tested against the number of training cases by undersampling Datasets 1 and 2B at different degrees, resulting into subsets of 20% to 80% of the original datasets. The sampling was based on a uniform random sampling scheme maintaining the equal class distribution of the original dataset. The results are illustrated in Fig. 2, where it can be noticed that the classification performance is degraded with a training set size that is less than 60% of the original one (significantly in the case of Dataset 2B). Considering that the computational complexity of the WCNN architecture depends on the size of the input images, undersampling of images was also investigated. However, the best classification performance was achieved with the original resolution images.

### *B. Comparisons with Conventional Supervised Classification Schemes*

The objective of this subsection is to compare the classification performance of WCNN with conventional, patch-based supervised classification schemes. The conventional patch-based supervised learning was evaluated using two different patch-based CNN methods and an SVM patch-based method. The first of these methods, can be considered as a baseline CNN (*b*-CNN) approach [18], whereas the second one, is the state-of-the-art approach proposed by Sekuboyina *et al* [27]. In the case of the *b*-CNN and SVM-based approaches, the raw RGB pixel values of the patches were used as inputs to the classifiers. RGB patches of 64×64 and 32×32 pixels were sampled from both Datasets 1 and 2B. The approach of Sekuboyina *et al* was applied on the same datasets using the *a*-channel of CIE-Lab with a patch size of 36×36 pixels for training, as suggested in [27]. The available pixel-level annotations of the datasets were used to characterize the patches as abnormal or normal. A patch was characterized as abnormal if the majority of its pixels were belonging to the abnormal class. These characterizations were used as binary training targets. A balanced set of 6,000 patches per class was used from Dataset 1, and another balanced set of 4,500 patches per class was used from Dataset 2B. The same number of patches was randomly selected regardless of the patch sizes. From each dataset, 2/3 of the dataset were used for training, and 1/3 was used for testing.

TABLE II
PERFORMANCE OF *b*-CNN, SEKUBOYINA *ET AL*, AND SVM FOR THE CLASSIFICATION OF IMAGE PATCHES FROM DATASETS 1 AND 2B (D1-D2B)

| Measure | *b*-CNN 64×64 | | *b*-CNN 32×32 | | Sekuboyina *et al* [27] | | SVM | |
|---|---|---|---|---|---|---|---|---|
| | D1 | D2B | D1 | D2B | D1 | D2B | D1 | D2B |
| Accuracy | **0.839** | **0.846** | 0.821 | 0.830 | 0.804 | 0.824 | 0.622 | 0.552 |
| Sensitivity | 0.872 | 0.908 | 0.852 | 0.840 | 0.817 | 0.840 | 0.986 | 0.767 |
| Specificity | 0.807 | 0.770 | 0.788 | 0.821 | 0.793 | 0.804 | 0.259 | 0.339 |

TABLE III
PERFORMANCE OF THE CONVENTIONAL CNN-BASED METHODS FOR THE CLASSIFICATION OF WHOLE IMAGES OF DATASETS 1 AND 2B (D1-D2B)

| Measure | WCNN | | *b*-CNN | | Sekuboyina *et al* [27] | |
|---|---|---|---|---|---|---|
| | D1 | D2B | D1 | D2B | D1 | D2B |
| Accuracy | **0.909** | **0.892** | 0.717 | 0.541 | 0.619 | 0.517 |
| Sensitivity | 0.930 | 0.924 | 0.961 | 0.610 | 0.957 | 0.952 |
| Specificity | 0.885 | 0.858 | 0.413 | 0.478 | 0.251 | 0.091 |

The classification of the patches using the *b*-CNN and the Sekuboyina *et al* approaches was implemented using the same CNN architectures and parameters as suggested in [18] and [27], respectively. Since the number of available patches was sufficiently larger than in [27] (where another, smaller KID dataset [25] was used), the methods used for artificial sample generation in that study were not applied.

The results obtained using the patch-based CNN architectures and the best SVM classifier (using 32×32-pixel patches and Gaussian kernel function) for the classification of the image patches are summarized in Table II, in terms of accuracy, sensitivity and specificity [59]. The performance of the SVM using 64×64-pixel patches was lower, resulting in an accuracy of approximately 0.5 for both datasets. The classifiers of both the *b*-CNN and the Sekuboyina *et al* outperform the SVM regardless the patch sizes. This result also demonstrates the enhanced feature extraction capability of the CNN methods over SVM. Also, the fact that all methods perform slightly better in the classification of Dataset 2B shows that the patterns of the anomalies in WCE images can be more easily discriminated than those of the anomalies in gastroscopy.

In order to compare the patch-based CNN methods with the WCNN, which classifies whole images, we extrapolated the classification results at image-level. This was possible by applying the following rule: a whole image is characterized as normal when all of its patches are classified as normal, whereas it is classified as abnormal if at least an abnormal patch is found in the image. Using different proportions of abnormal to normal patches per image this rule provides better sensitivity for all of the compared methods, since smaller

TABLE IV
10-FOLD CV CLASSIFICATION RESULTS OF THE WCNN AND STATE-OF-THE-ART WEAKLY SUPERVISED METHODS ON DATASETS 1-2 (D1-D2)

| Measure | WCNN | | Zhang *et al* [14] | | Jia and Meng [20] | | Yuan *et al* [40] | | Vasilakakis *et al* [19] | |
|---|---|---|---|---|---|---|---|---|---|---|
| | D1 | D2 | D1 | D2 | D1 | D2 | D1 | D2 | D1 | D2 |
| AUC | **0.963** | **0.814** | 0.951 | 0.773 | 0.902 | 0.705 | 0.940 | 0.709 | 0.946 | 0.802 |
| Accuracy | 0.899 | 0.775 | 0.851 | 0.760 | 0.827 | 0.690 | 0.867 | 0.696 | 0.892 | 0.768 |
| Sensitivity | 0.907 | 0.362 | 0.930 | 0.537 | 0.806 | 0.602 | 0.876 | 0.432 | 0.911 | 0.454 |
| Specificity | 0.882 | 0.913 | 0.779 | 0.836 | 0.857 | 0.785 | 0.854 | 0.820 | 0.872 | 0.886 |

anomalies are less likely to be missed (especially those with a size of the order of the patch size). The results obtained from this image-level comparison of the best performing patch-based *b*-CNN (64×64) and the Sekuboyina *et al* methods at pixel-level, are summarized in Table III. It is evident that the proposed WCNN outperforms both the compared patch-based CNN methods, although it is not trained with patches characterized as abnormal or normal based on pixel-level annotations. One can also observe in this table that the performance of all the methods on Dataset 2B is lower. This can be attributed mainly to the larger number of positives (lower specificity), which can be explained by the fact that the WCE dataset includes more anomalies of smaller size than those in the gastroscopy dataset. WCNN is less affected by the presence of smaller anomalies, since it exhibits a significantly higher specificity.

### C. Comparisons with Weakly Supervised Schemes

The previously described experiments were performed using the fixed training and testing sets, specified in Section IV. This was necessary for experimentation efficiency, since the first one (Section V.A) involved several repetitions of algorithm executions, and the second one (Section V.B) involved training with several sub-images.

Considering that weakly supervised methods use whole images for training and testing, a more thorough evaluation, using 10-fold Cross Validation (CV) was computationally feasible, using Datasets 1 and 2 (which is larger than the balanced Dataset 2B used in the previous experiments). This enables a less biased evaluation with respect to the selection of the training and testing sets, by randomly splitting the dataset into 10 non-overlapping parts. Out of the 10 parts, 9 were used for training and one for testing, repeatedly, until each part was used for testing once. The classification performance was investigated using Receiver Operating Characteristic (ROC) curves. An ROC curve depicts relative tradeoffs between benefits (correct decisions about abnormal cases, characterized as True Positives, TPs) and costs (false decisions about normal cases, characterized as False Positives, FPs) [59]. From a medical viewpoint, the ROC yields a pure measure of diagnostic accuracy, independent of the diagnostic criterion and of the frequencies of the alternative conditions under study [60]. With respect to the anomaly detection problem, the respective conditions are defined by the presence of an abnormal tissue within an endoscopic image or not. With respect to the localization problem, the respective conditions are defined as whether a point diagnosed as abnormal is located within an abnormal image area or not. The Area Under the ROC (AUC) is an overall summary measure of diagnostic accuracy [61]. In order to enable comparisons between the ROC curves, the AUC was used as a classification performance measure which, unlike accuracy, is relatively robust for datasets with imbalanced class distributions [62], as in the case of Dataset 2.

WCNN was compared with four state-of-the-art weakly supervised approaches reviewed in Section II. These include the CNN-based approaches of Zhang *et al* [14], and of Jia and Meng [20], and the BoW-based approaches of Yuan *et al* [40], and of Vasilakakis *et al* [19] using SVM as a classification scheme and the optimal parameters suggested in the respective studies. The average results obtained over the CV evaluation are summarized in Table IV and the standard deviation of the measurements was of the order of $10^{-2}$. Overall, focusing on the AUC measures, WCNN performs better than the compared weakly supervised schemes, with a significant advantage over Yuan's *et al* in the classification of Dataset 2. The classification performance of WCNN is almost equivalent to that of Zhang's *et al* method on Dataset 1 and to that of Vasilakakis' *et al* BoW-based approach for Dataset 2. It should be noted that results of previous studies in Dataset 1 have been presented in the context of the evaluation of the DSSVM method [24] (Section II). The reported AUC in that study was lower than that of the WCNN, reaching 0.898.

To make the comparison between the weakly supervised methods even more challenging, a third, larger and more diverse dataset was created by merging Dataset 1 and Dataset 2. This new dataset, referred to as Dataset 3, consists of a total of 3,050 images (698 images from Dataset 1 and 2,352 images from Dataset 2). Using the same 10-fold CV evaluation methodology used with Datasets 1 and 2, the classification performance of the proposed WCNN was again higher in terms of AUC on Dataset 3 reaching 0.861. The respective performance of Zhang's *et al* method [14] was 0.803, of Jia and Meng's method [20] was 0.846, of Yuan's *et al* method [40] was 0.681, and of Vasilakakis' *et al* method [19] was 0.840. The anomaly detection performances of all the compared methods on Dataset 3, except from the performance of Yuan *et al*, were higher than those obtained with Dataset 2 and lower than those obtained with Dataset 1. The underperformance of Yuan's *et al* method could be explained by the fact that it is based solely on texture features. Thus, the discrimination of anomalies with different color characteristics cannot be performed as effectively as with the other methods.

### D. Evaluation of Salient Point Detection

Following the classification of the GIE images, implemented by the best performing WCNN in the first phase of the proposed methodology, the second phase aims to detect

TABLE V
RESULTS OF SALIENT POINT DETECTION ALGORITHMS ON DATASETS 1 AND 2 (D1-D2), IN TERMS OF NUMBERS OF DETECTED POINTS (MIN-MAX).

| Salient Points per Image | DSD | | CSD [26] | |
|---|---|---|---|---|
| | D1 | D2 | D1 | D2 |
| Relevant (on GI anomalies) | **1-12** | **1-7** | 2-288 | 1-478 |
| Total | **1-18** | **1-10** | 195-349 | 548-683 |

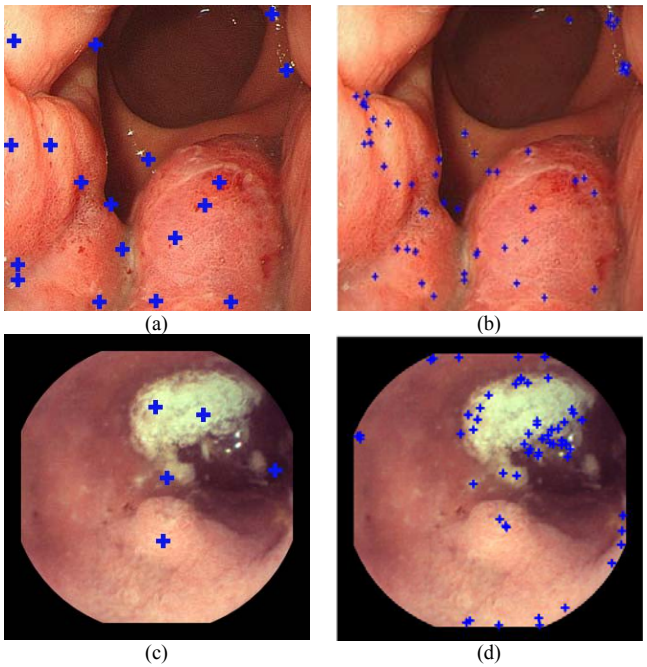


Figure 3. Salient point detection using DSD (left) and CSD [26] (right). (a-b) Images from Dataset 1. (c-d) Images from Dataset 2.

salient points, i.e., points that represent significant information with respect to the abnormal or normal classes. As described in Section III.B, this process requires a trained WCNN from which the saliency information is extracted from a deeper convolutional layer. In this study the third convolutional layer ($c$=3) has been selected for this purpose. We have visualized the feature maps from all layers and we observed that the first two convolutional layers of the network tend to encode less localized image features (in agreement with [49]). Also, the scale of the feature maps of deeper than the third layer is very small (≤42×42 pixels). This results in a very uncertain localization of the salient points in the original image, considering that the respective points in the input image are localized after linear upscaling to 320×320 pixels. In order to quantitatively validate these empirical observations, which led us to the selection of the third layer, an indicative test, inspired by the correlation-based feature section [63], was performed. For Datasets 1 and 2 we estimated Pearson's correlation coefficient between the pixel values of the maximal images derived from each convolutional layer, with the values of the binary masks derived from the ground truth pixel-level annotations of the anomalies. The binary masks are images of the same size with the input images, with pixel values 1 at the location of the anomalies and 0 elsewhere. For the estimation of the correlation they have been downscaled to match the dimensions of the maximal images. The correlation coefficients estimated from the first up to the last convolutional layer were 0.14, 0.05, 0.16, 0.13 and 0.04,

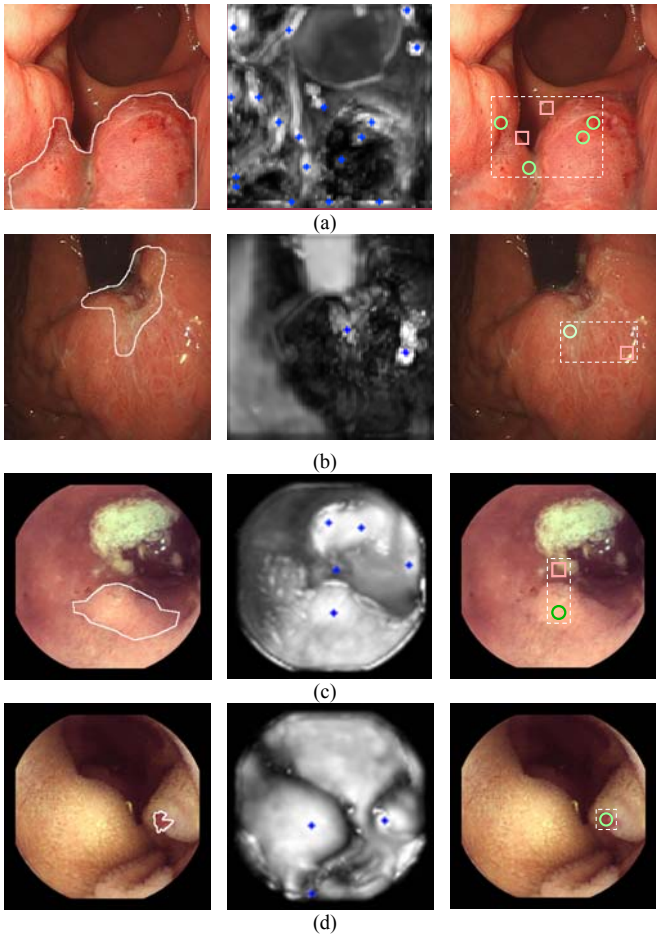


Figure 4. Lesion localization results on four abnormal images. The ground truth lesion areas are outlined on the original images presented in the left column. The maximal images are illustrated (scaled up) in the middle column, along with the respective salient points detected by DSD after Phase II. The localized lesions after the application of ICU in Phase III are presented in the right column. All points indicated with red square and green circular marks comprise the output of ICU. The green circular marks indicate the TPs and the red square marks indicate the FPs. (a-b) Images from Dataset 1. (c-d) Images from Dataset 2.

respectively. The highest value (0.16) observed for the third convolutional layer, is a quantitative indication that this layer was the most appropriate in this application context.

The salient point detection capability of DSD was evaluated on the test data of each fold of the 10-fold CV experiment described in the Section V.C, in terms of the relevant number of the salient points produced, i.e. the points belonging to GI anomalies. The results were compared to the Color Saliency Detector (CSD) proposed in [26]. The tolerance parameter ($t$) was empirically set to 120, and in the case of the CSD, the parameters were selected as suggested in [26].

The results obtained per image are presented in Table V. It can be noticed that the total number of points produced by DSD is significantly smaller and more relevant than the points produced by CSD. For example, in the case of Dataset 1 the maximum number of detected salient points by DSD in any abnormal image is 12, whereas CSD resulted in 288 points. This difference is significantly higher in the case of Dataset 2. Overall, using DSD, 46% of the salient points were relevant in

the case of Dataset 1, and 41% were relevant in the case of Dataset 2. Respecitvely, using CSD, 31% of the salient points obtained were relevant in the case of Dataset 1, and 16% were relevant in the case of Dataset 2. Both algorithms detected at least one relevant point per abnormal image. This is important as it simplifies the classification task performed in Phase III for the discrimination of the abnormal from the normal salient points, and consequently the localization of the GI anomalies. Indicative results from the application of DSD, in comparison with the results obtained using CSD [26], on representative images of the available datasets are illustrated in Fig. 3. Considering the respective ground truth annotations outlined in Figs. 4(a) and (c), Fig. 3 demonstrates the advantage of DSD to detect fewer and relevant salient points. Figure 4 includes additional examples of images from the available datasets as well as the respective maximal images with points detected by DSD. According to Table V the 18 points detected on the maximal image of Fig. 4(a), which corresponds to Fig. 3(a), is the maximum number of points detected in abnormal images in this dataset. The salient points detected by DSD in the rest of the images in Fig. 4(b-d) range between 2 and 5.

### *E. Evaluation of GI Anomalies Localization*

The performance of the proposed methodology in the localization of GI anomalies was evaluated on both Datasets 1 and 2, by extending the 10-fold CV scheme used in Section V.C to all the phases of the proposed methodology. In Phase III, ICU algorithm (Algorithm 2) filters the salient points detected in Phase II, by classification, and outputs a number of points that indicate possible locations of GI anomalies within abnormal images. The results obtained using the proposed PCFM features are compared to the results obtained using standard color features. To this end, a feature vector composed of the mean values of the respective CIE-Lab color space components ($a$, $b$) is used. The means are estimated over a 5×5 pixel neighborhood centered at the salient points. The choice of the window size used was determined as best, based on preliminary experimentation among window sizes of 1×1 to 16×16 pixels.

The number of clusters $Q$ and $R$ tested in the $k$-means algorithm varied from 2 to 10, and the number of $k$-means executions was $T$=10. The number of nearest neighbors tested was $K$=1,3,5,7 and the best performance was achieved by $K$=1. The results obtained from the output of ICU, in terms of AUC are illustrated in Fig. 5. This figure shows that best results in the two datasets are achieved for the least number of clusters ($k$=2). For higher values of $k$ the two classes are more difficult to discriminate. Using the PCFM features the best localization performance achieved in Dataset 1 is 0.848, and in Dataset 2 it is 0.877. Using CIE-Lab features the respective performances were 0.801 and 0.852. Overall, PCFM features perform better than CIE-Lab features, especially in the case of the larger dataset (D2).

Apart from these overall results, it is important to investigate the localization performance achieved at an image level. To this end the results per image were analyzed. This analysis showed that the average number of TP output points per image (i.e., points characterized as abnormal by the system that fall within the ground truth regions of the anomalies) in Dataset 1 was 1.78, ranging from 1 to a maximum of 7 points. The respective number of FPs (i.e., points that fall outside the ground truth regions of the anomalies but are characterized as abnormal by the system) was 0.74, ranging from 0 to 5. In Dataset 2 the average number of TPs was 1.14, ranging from 1 to 6. The average number of FPs was 0.69, ranging from 0 to 5 per image. Representative examples of results produced by the proposed methodology in Phase III are illustrated in the last column of Fig. 4. The output of ICU is a subset of the salient points detected by DSD, classified as abnormal. The dashed frame indicates the bounds of the region where these suspicious points are located. Points classified as normal are rejected. For example the result of ICU in Fig. 4(a) includes only two FP points (indicated with red squares), and four TP points. The FPs can be attributed to the lower illumination present in these regions. Each of the Figs. 4(b) and (c) has only one FP, and one TP. The FP of Fig. 4(b) corresponds to a

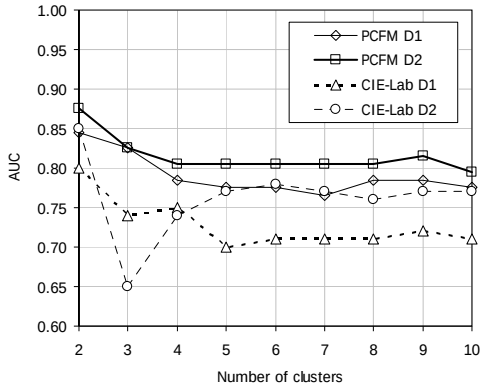


Figure 5. Anomaly localization performance of the proposed method using PCFM vs. CIE-Lab features on Datasets 1 and 2 (D1-D2), in terms of AUC, per target number of clusters $k$.

TABLE VI
ANOMALY LOCALIZATION RESULTS OVER ALL IMAGES OF DATASETS 1 AND 2 (D1-D2) USING ICU WITH PCFM FEATURES

| Detected Points per Image | D1 | | D2 | |
|---|---|---|---|---|
| | TP (%) | FP (%) | TP (%) | FP (%) |
| 0 | 0.0 | **54.3** | 0.0 | **52.9** |
| 1 | **59.9** | 28.0 | **92.9** | 31.0 |
| 2 | 22.4 | 11.2 | 5.3 | 9.6 |
| 3 | 11.2 | 3.7 | 0.6 | 3.9 |
| 4 | 4.7 | 1.9 | 0.0 | 2.2 |
| 5 | 0.9 | 0.9 | 0.6 | 0.4 |
| 6 | 0.0 | 0.0 | 0.6 | 0.0 |
| 7 | 0.9 | 0.0 | 0.0 | 0.0 |

TABLE VII
ANOMALY LOCALIZATION RESULTS OVER ALL IMAGES OF DATASETS 1 AND 2 (D1-D2) USING ENERGY MAPS

| Detected Points per Image | D1 | | D2 | |
|---|---|---|---|---|
| | TP (%) | FP (%) | TP (%) | FP (%) |
| 0 | 37.0 | 63.0 | 77.3 | 22.7 |
| 1 | 63.0 | 35.1 | 21.4 | 57.2 |
| 2 | 0.0 | 1.9 | 1.3 | 20.1 |
| 3 | 0.0 | 0.0 | 0.0 | 0.0 |
| 4 | 0.0 | 0.0 | 0.0 | 0.0 |
| 5 | 0.0 | 0.0 | 0.0 | 0.0 |
| 6 | 0.0 | 0.0 | 0.0 | 0.0 |
| 7 | 0.0 | 0.0 | 0.0 | 0.0 |

reflection, and the FP of Fig. 4(c) corresponds to a point of under-illuminated debris. Figure 4(d) includes a TP and it does not include any FP.

A summary of the localization results after the application of ICU using PCFM features, over all images of the available datasets is provided in Table VI. The output of this algorithm is a set of points (a subset of those detected by DSD) that are possibly abnormal (positive). This table lists the percentages of the images for which ICU produced 0, 1, 2,..., 7 TP and FP points (the maximum number of points per image was 7). For example, the percentage of images with 0 detected FP points per image (i.e., without any FP) was 54.3% in Dataset 1 and 52.9% in Dataset 2; the percentage of images with one TP point, was 59.9% in Dataset 1 and 92.9% in Dataset 2; and the percentage of images which had one FP point, was 28.0% in Dataset 1 and 31.0% in Dataset 2. It is notable that TPs were identified in all the abnormal images (the percentage of images with 0 detected TP points is 0.0%), and that the number of TPs or FPs per image did not exceed 7 in any case.

The performance of ICU was compared with the performance of a related state-of-the-art approach, which is based on the creation of energy maps [8]. According to that approach, the salient points detected by DSD are considered as 'fixations' or votes, and energy maps are created from this set of discrete fixations/votes. These fixation points are interpolated by a Gaussian function to build up the final energy map, from which the location of the global maximum of the saliency map is selected as the final output. After several experiments using Gaussian functions with different standard deviation values ($\sigma$ = 16, 32, 64), the best result, considering as a priority not to miss any anomalies, was obtained for $\sigma$ = 32. The respective percentages of images with TP and FP points are summarized in Table VII. It can be observed that the application of the energy-maps approach results in a significantly lower number of points per image; however, there are several images without any TP points detected (37.0% in Dataset 1 and 77.3% in Dataset 2). Thus, ICU is preferable.

Aiming to a further reduction of the FPs produced by ICU, the experiments were repeated with the energy maps used as a post-processing step that could possibly refine its output. However, although the FPs were reduced, the reduction of the TPs was unacceptable, as the percentage of images without any TP reached 43.7% in Dataset 1 and 77.2% in Dataset 2.

### F. *Broader Evaluation*

In order to demonstrate the broader usability of the proposed methodology, indicative experiments were performed on other datasets, using an already trained WCNN model. The model was trained on the entire Dataset 3 (Section V.C) which is composed of both Datasets 1 and 2. It was tested on a WCE video (named 'Case 1') of the KID database [45], and the colonoscopic images of the CVC-CLINIC and ETIS-LARIB databases [8].

*1) WCE Video:* The duration of the KID video is 2.8 hours and it was acquired with a 3 fps MiroCam CE. It contains several lesions, which according to 2 expert reviewers, include minuscule angiectasias (Fig.6a), lymphangiectasias (Fig.6b), and nodular lymphoid hyperplasias (Fig. 6c). The performance of the proposed methodology on the entire video is quantified by an AUC of 0.886 for anomaly detection, and an AUC of 0.769 for anomaly localization. The number of video segments containing sequences of lesions was 18, and in all of these segments at least one abnormal video frame was detected.

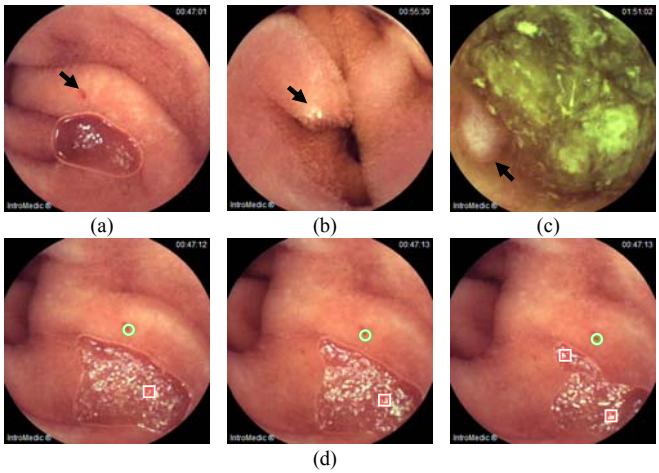


Figure 6. Video frames from 'Case 1' KID video sequence with anomalies. The arrows are used to indicate the location of each anomaly. (a) Angiectasia. (b) Lymphangiectasia. (c) Nodular lymphoid hyperplasia. (d) Indicative results from the localization of the angiectasia of Fig.6(a) in three consecutive frames randomly chosen from the entire sequence, which consists of 180 frames. All points represent the output of the proposed methodology, the green circular mark indicates a TP and the red square marks indicate FPs.

The video sequences usually contain consecutive frames with views of the same anomalies. Considering this, the temporal coherence of the response of the proposed methodology, with respect to the detection and localization of the lesions in *consecutive frames* was investigated. By following the evaluation approach suggested in [8], for each pair of consecutive frames with lesions, we calculated as a metric the percentage of these pairs in which the method provided correct output, i.e., the detection falls inside the region of the lesion, for both frames. On average, over all the frame sequences of the KID video containing lesions, this metric was estimated to be 50.4% with a standard deviation of 32.4%. In addition, the average *within sequence* detection stability was estimated to be 65.9% with a standard deviation of 28.1%. This metric represents the percentage of correct detections within the frames of each sequence. Considering that approx. 80% of the lesions were very small (area <256 pixels) these results are promising. A representative frame sequence with points indicating the possible lesion locations, as produced by the proposed methodology, is illustrated in Fig. 6(d). This sequence originates from the largest video segment where the anomaly was present in 180 frames. For that particular sequence the coherence in consecutive frames was 67% and the within sequence detection stability was 80%. The anomaly is generally well tracked and most of the FPs are due to highlights caused by the bubble present in the frames.

*2) Colonoscopic Images:* CVC-CLINIC database includes 612 Standard Definition (SD) frames and comprises 31 different polyps from 31 sequences; ETIS-LARIB database contains 196 High Definition (HD) frames and comprises 44 different polyps from 34 sequences. These datasets do not contain any normal frames; therefore, they cannot be considered as standalone training sets for our weakly supervised methodology. Considering that the polyp image datasets have been previously used for a comparative evaluation of relevant methodologies in [8], the results obtained are presented in terms of the same performance metrics used in that study. These include *precision* (Prec), *recall* (Rec), *F1* and *F2 measures*, estimated from the TP, FP, and False Negative (FN) cases, considering that the compared methods enable lesion localization. More specifically, if the output of the method is within the polyp region, the method is said to be providing a TP. Only one TP is considered per polyp, no matter how many detections fall within the polyp. Any detection that falls outside the polyp is considered a FP. The absence of alarm in images with a polyp is considered a FN, counting one per each polyp in the image that has not been detected. There are no images without polyps in these datasets; therefore, the number of True Negative (TN) cases was omitted.

All testing images were downscaled to 320×320 pixels, so as to fit the dimensions of the WCNN model (Fig. 1). The results obtained by the proposed methodology on the CVC-CLINIC and on the ETIS-LARIB databases are presented in Table VIII. This table also includes the results of the 3 top-ranked (based on the highest F1 measure) out of the 7 methods totally compared in [8]. However, it should be noted that these results were obtained in [8] by using the images of CVC-CLINIC database for training and the images of ETIS-LARIB databases for testing. No results were reported on CVC-CLINIC database in that study. All the results in Table VIII are ranked in a descending order by F1 measure. It can be noticed that the proposed methodology is ranked 3rd, although it was trained on a totally different dataset that does not include colonoscopic images of polyps (this justifies the overall lower localization performance as compared with the results presented in the previous subsection). The results obtained on the images of the CVC-CLINIC database are comparable.

The comparison performed focuses mainly on lesion localization, and we have followed it to provide directly comparable results with [8]. However the proposed methodology provides an intrinsic mechanism to detect abnormal frames in Phase I. This way, in ETIS-LARIB database it correctly detects 191 (97.5%) of the images as abnormal, and in the case of CVC-CLINIC database, it correctly detects 578 (94.5%) of the images as abnormal.

TABLE VIII
RESULTS ON POLYP IMAGE DATABASES.

| Method | TP | FP | FN | Prec* | Rec* | F1* | F2* |
|---|---|---|---|---|---|---|---|
| **ETIS-LARIB** | | | | | | | |
| Ranked-1 in [8] | 144 | 55 | 64 | 72.3 | 69.5 | 70.7 | 69.8 |
| Ranked-2 in [8] | 131 | 57 | 77 | 69.7 | 63 | 66.1 | 64.2 |
| **Proposed** | 94 | 69 | 114 | 57.7 | 45.2 | 50.7 | 47.2 |
| Ranked-3 in [8] | 110 | 226 | 98 | 32.7 | 52.8 | 40.4 | 47.1 |
| **CVC-CLINIC** | | | | | | | |
| **Proposed** | 284 | 223 | 362 | 56.0 | 44.0 | 49.3 | 45.9 |

* Percentage values (%).

TABLE IX
COMPARISON OF AVERAGE EXECUTION TIMES OF THE INVESTIGATED GI ANOMALY DETECTION METHODS ON DATASETS 1-2 (D1-D2)

| | D1 | | D2 | |
|---|---|---|---|---|
| | Training | Testing | Training | Testing |
| Images # | 628 | 70 | 2117 | 235 |
| Method | Time (min) | Time (min) | Time (min) | Time (min) |
| WCNN | 103 | **0.1** | 311 | **0.2** |
| Zhang *et al* [14] | **1** | **0.1** | **3** | **0.3** |
| Jia and Meng [20] | 70 | **0.1** | 72 | 0.4 |
| Yuan *et al* [40] | 7 | 0.6 | 24 | 2.4 |
| Vasilakakis *et al* [19] | 5 | 0.5 | 13 | 1.3 |
| *b*-CNN (64×64) | 4 | 0.3 | 4 | 0.8 |
| Sekuboyina *et al* [27] | 3 | 0.5 | 3 | 1.0 |

## G. Time-performance analysis

The performance of the proposed vs. the compared methods was evaluated also in terms of time-efficiency. All experiments were performed on a workstation with an Intel i5 2.5GHz CPU, 4GB RAM and an NVIDIA GeForce GTX 970 GPU. The CNN architectures were implemented using the Convolutional Architecture for Fast Feature Embedding (CAFFE) library [64], and the respective experiments were performed on the GPU. All other algorithms were implemented in MATLAB. The time-performance was measured in terms of average execution time per loop of the 10-fold CV process. The feature extraction times have been included in all cases for a fairer comparison with the CNN methods, which have an embedded feature extraction mechanism. The results obtained for the detection of abnormal images, i.e., the classification of the entire images as abnormal or normal (Phase I of the proposed methodology) are summarized in Table IX, with an error of ±0.05 min. The shorter training times regardless the dataset were achieved by Zhang *et al* [14]. In that method, CNN, which is pre-trained with non-medical images, is used only for feature extraction. This requires only a forward pass of the image data through the network. The classification of the feature vectors is implemented by an SVM, which is well-known for its efficiency in the training process [52]. However, it is also well-known that if the number of training vectors is large, SVMs tend to generate a large number of support vectors, which can slow down the testing process [65]. This explains the relatively longer testing times of Zhang *et al* and Jia and Meng [20] methods over the WCNN. The latter achieved the best time-performance during the testing phase, with a speedup over patch-based CNN methods reaching up to 5 for the larger dataset (D2). Both WCNN and Jia and Meng required longer training times than the rest of the compared methods. However, this does not affect their clinical usability since anomaly detection is based on already trained classification models. The shorter training time of Jia and Meng over the WCNN method can be explained by the less number of iterations required for convergence.

The respective times for the salient point detection and localization phases (Phases II and III) over the testing sets, were 0.07 min for Dataset 1 and 0.13 min for Dataset 2. For reference, the respective times of unsupervised segmentation algorithms that could be applied for both localization and size measurement of GI anomalies are: a) 69 min and 235 min using our recent approach for segmentation of various kinds of GI anomalies [45]; b) larger than 1.2 min and 3.9 min using the polyp segmentation method proposed in [9], since these are the times required only for the application of the *k*-means algorithm used in that method. Therefore, the advantage of using points instead of image segments for anomaly localization, with respect to time-performance, is significant.

## VI. Discussion and Conclusions

The methodology proposed in this paper aims to provide a cost-effective solution for automatic analysis of GIE video frame sequences, so as to enable both detection and localization of GI anomalies. This methodology is based on a WCNN, a generic neural network architecture that can be trained solely with semantically annotated images, indicating whether they contain anomalies or not. This is a great advantage over conventional, patch-based anomaly detection and localization approaches in GIE that require detailed annotation of training images, which is costly [3],[27].

The localization of GI anomalies has also been addressed with unsupervised image segmentation methods [9], [44], [45]. Such methods provide information about both the location and the area covered by an anomaly; therefore, they are also suitable for size measurement of GI anomalies. However, they are applicable only on images for which anomalies are present; otherwise by default they result in FP regions.

Pioneering studies on weakly-supervised classification methods for lesion detection in GIE have appeared in 2015 [39]. Considering the practical significance of this application the field has rapidly grown, with BoW and CNN-based approaches to play a protagonistic role (Section II). To date, only a few of the current methods cope with both the detection and localization problems [41], [42], [21]. These include methods that perform sufficiently, only if the weakly supervised learning is combined with detailed annotations at some level [41], [42]. In the context of GIE, only a preliminary study with a method tailored for the specific application of blood detection has been proposed [21].

To the best of our knowledge the methodology presented in our study is the first one coping with the general problem of GI anomaly localization using a WCNN. To achieve this, the WCNN-based classification of the GIE images is followed by the application of two novel algorithms: a) DSD algorithm which detects a few, but relevant salient points within the abnormal images; and b) ICU that refines the result of DSD by inferring the most suspicious of the salient points, using solely image-level information. This algorithm is based on clustering; however, unlike conventional approaches, it does not use clustering for image segmentation, and it does not exploit any pixel-level annotation. The output of this algorithm is a very small set of points that can attract the attention of a GIE video reviewer, so as to thoroughly examine the respective image locations. These points can be useful in a variety of ways, such as: seeds for other visualizations, e.g., using framing rectangles (Fig. 4), to indicate anomalous regions; sampling points for a secondary system capable of recognizing the types of the localized anomalies; and the initialization of contemporary lesion segmentation methods, as in [45], which are useful for size measurements but can prove time-inefficient for multi-frame video application.

Important outcomes that can be derived from this study about the proposed methodology include: a) it is both more effective and efficient than patch-based CNN-based approaches for detection of GI anomalies; b) although architecturally simpler than other state-of-the-art CNN-based approaches (e.g., approaches combining CNN with SVM classifiers [14], [20]), it performs better or equal to state-of-the-art weakly supervised approaches, while its advantages become more apparent with larger and more diverse datasets; c) with the use of DSD and ICU algorithms the GI anomalies are not only detected but also localized in both an effective and efficient way.

The proposed methodology was challenged to detect and localize anomalies in totally new datasets, including an entire WCE video with various anomalies, and colonoscopic images with polyps. The anomalies of the WCE video were vascular and polypoid lesions. Dataset 2 included such types of lesions. Neither Dataset 1 nor 2 included colon polyps. In both cases the proposed methodology performed sufficiently, considering the performances reported in [8], and the fact that the scope of our methodology extends to the detection and localization of various anomalies beyond polyps. These results indicate its broader usability, with a potential for improvement by using larger, even more diverse training sets.

Future research directions include the investigation of novel approaches to further improve the performance of the proposed methodology, e.g., by coping with intestinal content, which has been identified as a source of FPs. They also include the investigation of WCNN training algorithms with less computational requirements, alternative weakly-supervised approaches, experimentation on even larger datasets that could be of higher resolution, systematic evaluation on entire endoscopy videos (considering coherence and other issues identified in recent studies [8], [66]), evaluation of various cluster distance measures, and adaptation of the proposed methodology for application to other imaging domains.

## References


[1] A. Koulaouzidis, D. K. Iakovidis, A. Karargyris, and J. N. Plevris, "Optimizing lesion detection in small-bowel capsule endoscopy: from present problems to future solutions," *Expert review of gastroenterology & hepatology*, vol. 9, no. 2, pp. 217–235, 2015.

[2] D. K. Iakovidis and A. Koulaouzidis, "Software for enhanced video capsule endoscopy: challenges for essential progress," *Nature Reviews Gastroenterology & Hepatology*, vol. 12, no. 3, pp. 172–186, 2015.

[3] S. A. Karkanis, D. K. Iakovidis, D. E. Maroulis, D. A. Karras, and M. Tzivras, "Computer-aided tumor detection in endoscopic video using color wavelet features," *IEEE transactions on information technology in biomedicine*, vol. 7, no. 3, pp. 141–152, 2003.

[4] Y. Zheng, L. Hawkins, J. Wolff, O. Goloubeva, and E. Goldberg, "Detection of lesions during capsule endoscopy: physician performance is disappointing," *The American journal of gastroenterology*, vol. 107, no. 4, pp. 554–560, 2012.

[5] S. Beg, A. Parra-Blanco, and K. Ragunath, "Optimising the performance and interpretation of small bowel capsule endoscopy," *Frontline Gastroenterology*, p. flgastro–2017, 2017.

[6] M. Hoai, L. Torresani, F. D. la Torre, and C. Rother, "Learning discriminative localization from weakly labeled data," *Pattern Recognition*, vol. 47, no. 3, pp. 1523–1534, 2014.

[7] M. Liedlgruber and A. Uhl, "Computer-aided decision support systems for endoscopy in the gastrointestinal tract: A review," *Biomedical Engineering, IEEE Reviews in*, vol. 4, pp. 73–88, 2011.

[8] J. Bernal, N. Tajkbaksh, F. J. Sanchez, B. J. Matuszewski, H. Chen, L. Yu, Q. Angermann, O. Romain, B. Rustad, I. Balasingham, K. Pogorelov, S. Choi, Q. Debard, L. Maier-Hein, S. Speidel, D. Stoyanov, P. Brandao, H. Cordova, C. Sanchez-Montes, S. R. Gurudu, G. Fernandez-Esparrach, X. Dray, J. Liang, and A. Histace, "Comparative Validation of Polyp Detection Methods in Video Colonoscopy: Results From the MICCAI 2015 Endoscopic Vision Challenge," *IEEE Transactions on Medical Imaging*, vol. 36, no. 6, pp. 1231–1249, Jun. 2017.

[9] Y. Jia, "Polyps auto-detection in wireless capsule endoscopy images using improved method based on image segmentation," in *Robotics and Biomimetics (ROBIO), 2015 IEEE International Conference on*, 2015, pp. 1631–1636.

[10] A. V. Mamonov, I. N. Figueiredo, P. N. Figueiredo, and Y.-H. R. Tsai, "Automated polyp detection in colon capsule endoscopy," *IEEE transactions on medical imaging*, vol. 33, no. 7, pp. 1488–1502, 2014.

[11] E. Ribeiro, A. Uhl, and M. Häfner, "Colonic polyp classification with convolutional neural networks," in *Computer-Based Medical Systems (CBMS), 2016 IEEE 29th International Symposium on*, 2016, pp. 253–258.

[12] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, 2015, pp. 79–83.

[13] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, May 2016.

[14] R. Zhang, Y. Zheng, T. W. C. Mak, R. Yu, S. H. Wong, J. Y. W. Lau, and C. C. Y. Poon, "Automatic Detection and Classification of Colorectal Polyps by Transferring Low-Level CNN Features From Nonmedical Domain," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 1, pp. 41–47, Jan. 2017.

[15] A. Karargyris and N. Bourbakis, "Detection of small bowel polyps and ulcers in wireless capsule endoscopy videos," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 10, pp. 2777–2786, 2011.

[16] F. Van Der Sommen, S. Zinger, E. J. Schoon, and P. de With, "Supportive automatic annotation of early esophageal cancer using local gabor and color features," *Neurocomputing*, vol. 144, pp. 92–106, 2014.

[17] G. Wimmer, A. Vécsei, and A. Uhl, "CNN transfer learning for the automated diagnosis of celiac disease," in *Image Processing Theory Tools and Applications (IPTA), 2016 6th International Conference on*, 2016, pp. 1–6.

[18] S. V. Georgakopoulos, D. K. Iakovidis, M. Vasilakakis, V. P. Plagianakos, and A. Koulaouzidis, "Weakly-supervised Convolutional learning for detection of inflammatory gastrointestinal lesions," in *Imaging Systems and Techniques (IST), 2016 IEEE International Conference on*, 2016, pp. 510–514.

[19] M. Vasilakakis, D. K. Iakovidis, E. Spyrou, and A. Koulaouzidis, "Weakly-Supervised Lesion Detection in Video Capsule Endoscopy based on a Bag-of-Colour Features Model," *T. Peters et al. (Eds.): CARE 2016, LNCS 10170*, pp. 1–8, 2017.

[20] X. Jia and M. Q. H. Meng, "A deep convolutional neural network for bleeding detection in Wireless Capsule Endoscopy images," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2016, pp. 639–642.

[21] X. Jia and M. Q.-H. Meng, "A study on automated segmentation of blood regions in Wireless Capsule Endoscopy images using fully convolutional networks," in *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*, 2017, pp. 179–182.

[22] Y. Yuan, B. Li, and M. Q.-H. Meng, "Bleeding frame and region detection in the wireless capsule endoscopy video," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 2, pp. 624–630, 2016.

[23] Y. Cong, S. Wang, B. Fan, Y. Yang, and H. Yu, "UDSFS: unsupervised deep sparse feature selection," *Neurocomputing*, vol. 196, pp. 150–158, 2016.

[24] Y. Cong, S. Wang, J. Liu, J. Cao, Y. Yang, and J. Luo, "Deep sparse feature selection for computer aided endoscopy diagnosis," *Pattern Recognition*, vol. 48, no. 3, pp. 907–917, 2015.

[25] D. K. Iakovidis and A. Koulaouzidis, "Automatic lesion detection in capsule endoscopy based on color saliency: closer to an essential adjunct for reviewing software," *Gastrointestinal Endoscopy*, vol. 80, no. 5, pp. 877–883, 2014.

[26] D. K. Iakovidis and A. Koulaouzidis, "Automatic lesion detection in wireless capsule endoscopy□A simple solution for a complex problem," in *Image Processing (ICIP), 2014 IEEE International Conference on*, 2014, pp. 2236–2240.

[27] A. K. Sekuboyina, S. T. Devarakonda, and C. S. Seelamantula, "A convolutional neural network approach for abnormality detection in Wireless Capsule Endoscopy," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, 2017, pp. 1057–1060.

[28] R. Zhu, R. Zhang, and D. Xue, "Lesion detection of endoscopy images based on convolutional neural network features," in *Image and Signal Processing (CISP), 2015 8th International Congress on*, 2015, pp. 372–376.

[29] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," in *Proceedings of the IEEE*, 1998, vol. 86, no. 11, pp. 2278–2324.

[30] H. Greenspan, B. van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1153–1159, 2016.

[31] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural Network," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1207–1216, May 2016.

[32] M. J. J. P. van Grinsven, B. van Ginneken, C. B. Hoyng, T. Theelen, and C. I. Sánchez, "Fast Convolutional Neural Network Training Using Selective Data Sampling: Application to Hemorrhage Detection in Color Fundus Images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1273–1284, May 2016.

[33] Y. Huang, H. Zheng, C. Liu, X. Ding, and G. Rohde, "Epithelium-stroma classification via convolutional neural networks and unsupervised domain adaptation in histopathological images," *IEEE Journal of Biomedical and Health Informatics*, vol. PP, no. 99, pp. 1–1, 2017.

[34] H. R. Roth, L. Lu, J. Liu, J. Yao, A. Seff, K. M. Cherry, L. Kim, and R. M. Summers, "Improving Computer-Aided Detection Using Convolutional Neural Networks and Random View Aggregation," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1170–1181, 2016.

[35] G. Carneiro, T. Peng, C. Bayer, and N. Navab, "Automatic Quantification of Tumour Hypoxia From Multi-Modal Microscopy Images Using Weakly-Supervised Learning Methods," *IEEE Transactions on Medical Imaging*, vol. 36, no. 7, pp. 1405–1417, Jul. 2017.

[36] H. Chen, X. Wu, G. Tao, and Q. Peng, "Automatic content understanding with cascaded spatial–temporal deep framework for capsule endoscopy videos," *Neurocomputing*, vol. 229, pp. 77–87, 2017.

[37] S. Segui, M. Drozdzal, G. Pascual, P. Radeva, C. Malagelada, F. Azpiroz, and J. Vitria, "Generic feature learning for wireless capsule endoscopy analysis," *Computers in biology and medicine*, vol. 79, pp. 163–172, 2016.

[38] S. Wang, Y. Cong, H. Fan, L. Liu, X. Li, Y. Yang, Y. Tang, H. Zhao, and H. Yu, "Computer-Aided Endoscopic Diagnosis Without Human-Specific Labeling," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 11, pp. 2347–2358, 2016.

[39] S. Wang, Y. Cong, H. Fan, Y. Yang, Y. Tang, and H. Zhao, "Computer aided endoscope diagnosis via weakly labeled data mining," in *Image Processing (ICIP), 2015 IEEE International Conference on*, 2015, pp. 3072–3076.

[40] Y. Yuan, B. Li, and M. Q.-H. Meng, "Improved bag of feature for automatic polyp detection in wireless capsule endoscopy images," *IEEE Transactions on Automation Science and Engineering*, vol. 13, no. 2, pp. 529–535, 2016.

[41] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free? – Weakly-supervised learning with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 685–694.

[42] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1742–1750.

[43] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *International Conference on Learning Representations (ICLR)*, 2015.

[44] M. Ganz, X. Yang, and G. Slabaugh, "Automatic segmentation of polyps in colonoscopic narrow-band imaging data," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 8, pp. 2144–2151, 2012.

[45] A. Koulaouzidis, D. K. Iakovidis, D. E. Yung, E. Rondonotti, U. Kopylov, J. N. Plevris, E. Toth, A. Eliakim, G. W. Johansson, W. Marlicz, and others, "KID

Project: an internet-based digital video atlas of capsule endoscopy for research purposes," *Endoscopy International Open*, vol. 5, no. 06, pp. E477–E483, 2017.
[46] R. Qian, Y. Yue, F. Coenen, and B. Zhang, "Visual attribute classification using feature selection and convolutional neural network," in *2016 IEEE 13th International Conference on Signal Processing (ICSP)*, 2016, pp. 649–653.
[47] T. Rui, J. Zou, Y. Zhou, J. Fei, and C. Yang, "Convolutional neural network feature maps selection based on LDA," *Multimedia Tools and Applications*, Apr. 2017.
[48] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio, "Object recognition with gradient-based learning," *Shape, contour and grouping in computer vision*, pp. 823–823, 1999.
[49] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*, 2014, pp. 818–833.
[50] J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, and others, "Fiji: an open-source platform for biological-image analysis," *Nature methods*, vol. 9, no. 7, pp. 676–682, 2012.
[51] P. Soille, *Morphological image analysis: principles and applications*. Springer Science & Business Media, 2013.
[52] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. Elsevier/Academic Press, 2008.
[53] S. K. Tasoulis, D. K. Tasoulis, and V. P. Plagianakos, "Random direction divisive clustering," *Pattern Recognition Letters*, vol. 34, no. 2, pp. 131–139, 2013.
[54] L. Bottou, "On-line Learning in Neural Networks," D. Saad, Ed. New York, NY, USA: Cambridge University Press, 1998, pp. 9–42.
[55] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *The International Conference on Learning Representations (ICLR)*, 2015.
[56] S. J. Reddi, S. Kale, and S. Kumar, "On the Convergence of Adam and Beyond," in *International Conference on Learning Representations (ICLR)*, 2018.
[57] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 3, no. 52, 1965.
[58] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, Dec. 1945.
[59] T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.
[60] J. A. Swets, "ROC analysis applied to the evaluation of medical imaging techniques.," *Investigative radiology*, vol. 14, no. 2, pp. 109–121, 1979.
[61] D. Alemayehu and K. H. Zou, "Applications of ROC analysis in medical research: recent developments and future directions," *Academic radiology*, vol. 19, no. 12, pp. 1457–1464, 2012.
[62] F. J. Provost and T. Fawcett, "Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions.," in *KDD*, 1997, vol. 97, pp. 43–48.
[63] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.
[64] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional Architecture for Fast Feature Embedding," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, 2014, pp. 675–678.
[65] C. J. Burges, "Simplified support vector decision rules," in *ICML*, 1996, vol. 96, pp. 71–77.
[66] 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. CARE 2017, CLIP 2017. Lecture Notes in Computer Science*, Springer, Cham, 2017, pp. 29–41.