IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 13 July 2022, accepted 15 August 2022, date of publication 25 August 2022, date of current version 6 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3201515

RESEARCH ARTICLE

Multichannel Residual Cues for Fine-Grained Classification in Wireless Capsule Endoscopy

ANUJA VATS[1], KIRAN RAJA[1], MARIUS PEDERSEN[1], (Member, IEEE), AND AHMED MOHAMMED[1,2], (Member, IEEE)

[1]Department of Computer Science, NTNU, 2815 Gjøvik, Norway
[2]SINTEF Digital, 0373 Oslo, Norway

Corresponding author: Anuja Vats (anuja.vats@ntnu.no)

This work was supported by the Research Council of Norway under Grant 300031.

ABSTRACT Early diagnosis of gastrointestinal pathologies leads to timely medical intervention and prevents disease development. Wireless Capsule Endoscopy (WCE) is used as a non-invasive alternative for gastrointestinal examination. WCE can capture images despite the structural complexity presented by human anatomy and can help in detecting pathologies. However, despite recent progress in fine-grained pathology classification and detection, limited works focus on generalization. We propose a multi-channel encoder-decoder network for learning a generalizable fine-grained pathology classifier. Specifically, we propose to use structural residual cues to explicitly impose the network to learn pathology traces. While residuals are extracted using well-established 2D wavelet decomposition, we also propose to use colour channels to learn discriminative cues in WCE images (like red color in bleeding). With less than 40% data (fewer than 2500 labels) used for training, we demonstrate the effectiveness of our approach in classifying different pathologies on two different WCE datasets (different capsule modalities). With a comprehensive benchmark for WCE abnormality and multi-class classification, we illustrate the generalizability of the proposed approach on both datasets, where our results perform better than the state-of-the-art with much fewer labels in abnormality sensitivity on several of nine different pathologies and establish a new benchmark with specificity $> 97\%$ across classes.

INDEX TERMS Pathology classification, wireless capsule endoscopy, fine grained, residual cues.

I. INTRODUCTION

The gastrointestinal tract, which plays a vital role in absorbing essential nutrients, also exhibits susceptibility to various intestinal pathologies such as ulcers, polyps, lesions and inflammation, amongst others. The lack of proper management of inflammatory bowel disease can lead to encumbering economies with substantial costs at different stages of disease management pipeline [2] in addition to reducing the quality of life among patients [3], [4]. Therefore, the early diagnosis of such pathological occurrences as colorectal polyps or obscure intestinal bleeding can lead to timely intervention and possible prevention of further disease development to extreme cases such as inflammatory bowel diseases and cancers [5], [6], [7]. While standard diagnostic approaches like colonoscopy and gastroscopy enable visualization of the upper and lower gastrointestinal tract, anatomical challenges prevent satisfactory examination of the small bowel through these techniques [8]. As a consequence, Wireless Capsule Endoscopy (WCE) has become an increasingly preferred non-invasive alternative, especially for small bowel examinations [9].

Despite progress in WCE imaging [10], a medical practitioner needs to spend reasonable time analysing frames in excess of 60,000 [11], amounting approximately to 45 to 60 minutes [12] of the practitioner's time for large bowel diagnosis. Despite the substantial time requirement, the diagnosis has susceptibility to human error [13]. Reliable computer-aided classification tools for detecting normal from pathological conditions can be paramount in cutting down individual diagnostic times and associated costs [14]. A number of approaches have been proposed in the last

The associate editor coordinating the review of this manuscript and approving it for publication was M. Sabarimalai Manikandan.

FIGURE 1. An example illustration of pathological images from capsule endoscopy with different cases of pathologies from Kvasir-Capsule dataset (referred as D1 in this work).

decade to classify the images/videos using classical techniques as well as deep learning based architectures [15]. While classical approaches are carefully engineered by understanding the structure and appearance of pathology in images [16], deep learning architectures identify and classify pathological appearances by learning to extract task-beneficial features from large datasets.

Although classifying a normal image against pathology is of great clinical significance, there are limited works that attempt to do fine-grained classification involving different pathologies within a single framework [8], [17], leading to a problem of generalization in real-time. The dearth of works in this direction can be attributed to limitations pertaining to the private nature of this data and the scarcity of sufficiently varied examples of different cases, often leading to inadequately representative, unbalanced datasets for one or more pathologies. While machine learning-based algorithms enable prediction of normal against pathological conditions with high confidence [18], [19], [20], attempts at discriminating between different pathological conditions presents a considerable generalization challenge, even with large datasets, as also noted in the multi-centric study by Ding *et al.* [8]. Figure 1 shows few examples of pathological images obtained from WCE from the Kvasir Capsule Dataset [1] illustrating the varying appearances of pathologies in different images and the challenge in classifying them with high confidence, especially with considerations to changing anatomy within the gastrointestinal tract.

We present a new approach for classifying various pathologies from normal images using a multi-channel Convolutional Neural Network (CNN) based cascaded encoder-decoder network with a ResNet [21] backbone. We observe that occurrence of pathology is coupled with tissue transitions in different directions and that these transitions could provide useful cues for pathology identification. In line with this, we propose to extract directional residuals based on 2D-Wavelet decomposition. We illustrate this assertion with a toy example in Figure 2 where an image is decomposed into three directions that capture different directional aspects of the information contained within the original image. Concretely, the original image (top-left) is decomposed in horizontal (top-right), vertical (bottom-left) and diagonal (bottom-right) directions using 2D-Discrete wavelet transform. This directional information for each orientation is then subtracted from the original image to provide the corresponding residuals, which we refer to as structural residual cues. We experimentally verify that in WCE, these directional residuals not only capture the directional transitions resulting from an abnormality but carry additional structural details particular to different abnormalities that allows learning better discrimination between normal and pathological cases (Table. 1) as well as between different pathological cases (Table. 4).

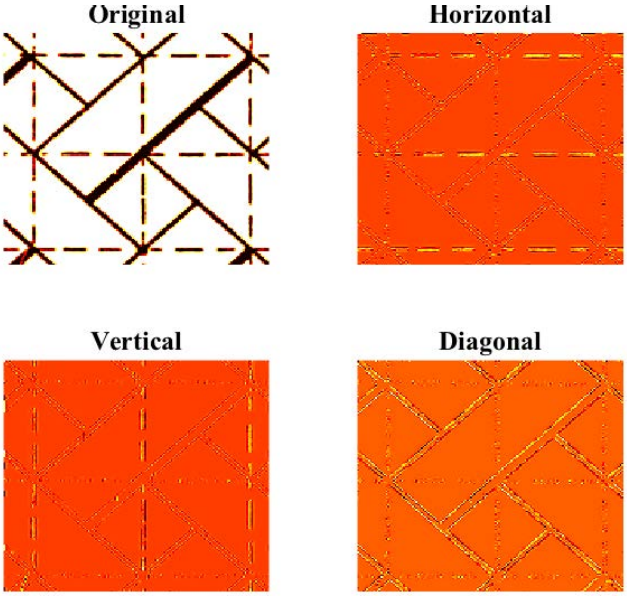

FIGURE 2. A toy example illustrating the DWT decomposition where the structural residuals can be seen along the horizontal, vertical and diagonal directions.

In addition to structural changes, transitions in colour across different spatial regions of an image can provide further clues that medical practitioners often use for deciding on the presence of a pathological condition. For instance, the red colour is pronounced when a blood clot is observed in the tract compared to an ulcer which appears yellowish. Motivated by such observations, we use colour channels in addition to structural residuals to learn the classification model robustly. Thus, our approach works by employing six different channels consisting of both structural residuals and appearance factors in different colour spaces.

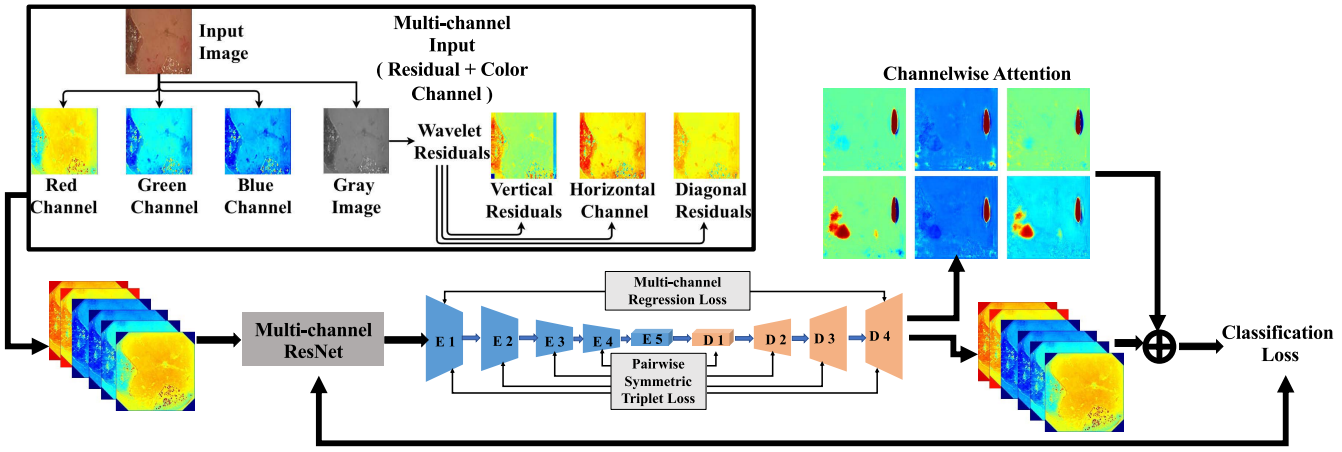

FIGURE 3. Proposed architecture for multi-channel CNN with cascaded Encoder-Decoders. The images are shown in colour-map for better visualization.

Finally, we propose to exploit intermediate representations at different stages of the cascaded Encoder-Decoder network towards obtaining a final decision to fully utilize features of different scales and sensitivity. We do this through two ways: (a) enforcing a symmetrical loss between different intermediate latent representations between the encoder and the decoder, using a standard triplet loss at the batch level. (b) incorporating a pixel-wise reconstruction loss between the ends, i.e., the input and the decoder output. In addition, a standard Negative Log-Likelihood loss is employed to train the network in a supervised manner to classify the normal versus pathology images by taking full advantage of the 6 channel CNN.

We validate our proposed approach on the fine-grained classification of two WCE datasets - Kvasir-Capsule dataset (D1) [1] and the Computer Assisted Diagnosis for Capsule Endoscopy Database (CAD-CAP) from the Giana Endoscopic Vision Challenge (D2) [22], with up to nine different pathology categories. The obtained results surpass previous baselines in per-pathology sensitivity while closing the gap in sensitivity and specificity with other works with significantly higher labels used in training. We demonstrate the applicability of the proposed approach for classifying many different pathologies (using the same network, no additional architectural changes) from normal across two datasets. In addition, we evaluate the strength of features in characterizing challenging categories through a multi-class classification between pathology categories. We present a detailed analysis on the explainability of the network through analysis of channel contributions. The main contributions are:

- We propose a unified multi-channel CNN based cascaded Encoder-Decoder network to classify multiple WCE pathologies with consistently high sensitivity and specificity.
- We demonstrate generalizability in challenging cross-pathology (nine different non-overlapping classes of datasets) and cross-dataset settings (with different capsule modalities).
- Unlike previous works in WCE, our work presents a more comprehensive baseline for evaluating future works, with the inclusion of multiple commonly occurring pathologies in WCE.

In the rest of the paper, we present the rationale and proposed approach in detail in Section II, followed by a summary of the datasets employed for validation. Section IV presents the detailed analysis of the results and Section V provides the discussion on explainability. Further, Section VI provides the results of the ablation studies with respect to the losses and Section VIII discusses the conclusions.

II. METHODOLOGY

The image of a healthy organ typically exhibits different visual properties from the one with diseases/pathology as illustrated in Figure 1. A pathological image I_p can be asserted as a corrupted version of a healthy image I_h where the corruption refers to the specific structural irregularities over and above normal variations, resulting from the pathology. Thus, a pathology image I_p, can be described as a some function f of healthy image I_h and corruption $\omega(\cdot)$ i.e, $I_p = f(I_h) + \omega(\cdot)$ (the corruption being a complex interplay of patient related pathological and physiological variables). Assuming that the healthy images are abundant, we formulate pathological characteristics as outliers deviating from normal healthy images with different abnormal appearances. Thus, we propose an approach based on such an assertion to learn the outliers in the presence of pathology as detailed in the upcoming section.

A. MULTICHANNEL INPUT THROUGH STRUCTURAL RESIDUAL CUES AND COLOUR INFORMATION

The pathological and normal images may appear very similar in multiple cases, given the complex nature of these images. However, some superficial deposits or patterns can be observed in images with pathology when captured with WCE, and using only RGB images may not necessarily lead to features robust enough for learning optimal

classifiers often leading to mis-classifications. Therefore, we first seek to obtain crucial discriminative differences. As the pathology characteristics can appear in all possible orientations and shapes, it is vital to extract pathology traces in all such directions. Although such information can be extracted using different spatial approaches, we rely upon the 2D-discrete wavelets transforms (2D-DWT) to capture details in horizontal, vertical, and diagonal directions as illustrated in Figure 2 to obtain the traces of pathology in different orientations. The wavelets from each of the horizontal, vertical, and diagonal directions provide clear transitions between the areas of pathology and non-pathological regions. It has to be, however, noted that such transitions can also appear due to the tissues that are normally present without any pathology. To deal with such a situation, we take the residual differences between the original image and each of the responses from horizontal, vertical, and diagonal directions leading to a set of structural residuals. We first obtain the gray-level image I_g of I and then obtain the residuals along horizontal, vertical and diagonal directions I_h^r, I_v^r, I_d^r, represented as

$$\begin{aligned} I_h^r &= I_g - \psi^{(h)} I_g \\ I_v^r &= I_g - \psi^{(v)} I_g \\ I_d^r &= I_g - \psi^{(d)} I_g \end{aligned}$$

where ψ represents the 2D-DWT in different directions at level 1.

While all residual images capture structural transitions in different orientations and discard colour information, our approach may suffer from missing colour information. To account for this, we combine the structural residual information with the colour channel information to create a multi-channel input. Thus, the RGB channels of input image I is appended to form a final six channel input resulting in $\{I_h^r, I_v^r, I_d^r, I_R, I_G, I_B\}$ for learning the classifier.

B. LEARNING TO CLASSIFY RESIDUAL CUES

As discussed earlier, a pathological image can be considered as a corrupted variant of the healthy image given by $I_p = f(I_h) + \omega(\cdot)$. Thus, we assert that learning the cues corresponding to $\omega(\cdot)$ leads to a better classification of the pathological images from healthy images as well as within themselves. We, therefore, propose an Encoder-Decoder architecture to learn the cues more effectively discriminating healthy images from pathology images.

Given the multi-channel input, we extract the features for learning the classifier using a Residual Network (ResNet-18) [21] backbone with 18 layers. Unlike the traditional three-channel input for a deep neural network, the input in our case has six channels, thus we modify the ResNet-18 architecture to process multi-channel inputs for feature extraction. Furthermore, to account for the limited data settings which can result in an overfitted network when learning from scratch, especially in our case for pathological images where the data scarcity cannot be overruled, we use the pre-trained weights learnt on ImageNet [23] for the last three channels (i.e., RGB) and set the weights of other three channels corresponding to residual differences to zero. To make the learning robust and incorporate all intermediate features, we adopt a U-Net [24] like Encoder-Decoder architecture, where we choose four Encoder (E) residual blocks followed by a corresponding number of Decoder (D) residual blocks. Such an Encoder-Decoder architecture helps us to decode the information back to generate the original six channel input. The feature maps from each layer are upsampled by a nearest-neighbour interpolation through a 2×2 convolution in each of the successive decoder blocks. We exploit the intermediate symmetrically encoded-decoded representations (i.e., Encoder-1 with Decoder-4) by using it to measure the differences using a standard triplet loss by formulating it as an anomaly problem inspired by a recent work by Feng *et al.* [25].

Learned feature representations for images of the same class must ideally lie in a compact sphere, while features corresponding to samples from dissimilar classes lie further away in the representation space. In an analogy, the representation of a pathological image must be such that it drifts away (exhibits smaller pairwise similarity) from the center of the cluster of healthy images in the feature space. If the input space is represented by ($\mathcal{X} \subseteq \mathbb{R}^d$) and the output space by ($\mathcal{Z} \subseteq \mathbb{R}^p$), a neural network with L hidden layers and corresponding set of weights $\mathcal{W} = \{\mathcal{W}^1, \ldots, \mathcal{W}^L\}$ learns a mapping between the input space and the output space as $\phi(\cdot; \mathcal{W}) : \mathcal{X} \rightarrow \mathcal{Z}$ such that $\mathcal{Z}$ exhibits the desirable properties described where $\mathcal{Z}$ is centered on a predetermined point c.

Given n healthy samples ($I_1^h, \ldots, I_n^h \subseteq \mathcal{X}$) and k pathological samples ($I_1^p, \ldots, I_k^p \subseteq \mathcal{X}$), the point c in the output space $\mathcal{Z}$ should lie close to healthy samples while the pathological samples are encoded such that they lie away from this healthy cluster.

$$\min_{\mathcal{W}} \frac{1}{h} \sum_{i=1}^{h} \left\| \phi(I_i^h; \mathcal{W}) - c \right\|^2 \tag{1}$$

$$\max_{\mathcal{W}} \frac{1}{p} \sum_{i=1}^{k} \left\| \phi(I_i^p; \mathcal{W}) - c \right\|^2 \tag{2}$$

Minimizing the distance for samples from healthy class while maximizing the distance for samples from pathological class from c can lead to better classification. To achieve this, we employ triplet loss L_t to minimize intra-class separability and maximize inter-class separability at the feature level [26]. Using all intermediate feature representations from the last encoder block to the last decoder block, we create triplets of healthy images and pathological images in each batch by randomly choosing healthy samples as anchors and the rest of the healthy samples as positives and pathological samples as negative. Thus, by enforcing a triplet loss that uses 6 input channels (residual and RGB), the loss estimation can be fully used to benefit from learning class separation.

TABLE 1. Comparison of performance between different datasets and approaches for abnormality detection. All pathologies are clubbed as a single class of abnormality and classified against the normal class.

Approach	Total no. of images	Sensitivity	Specificity	AUC	Accuracy	Abnormalities
Ding et. al. [8]	158 235	99.9	100	–	–	Inflammation,Polyps, bleeding, vascular disease, protruding lesion, diverticulum, parasite, ulcer, lymphatic follicular hyperplasia, lymphangiectasia Vs Normal
Iakovidis et. al [27]	1196	92.4	85.8	–	89.2	Vasular anomalies, polypoid anomalies inflammatory anomalies Vs Normal
Ours [D1]	5489	99.3	99.6	99.4	99.42	Angiectasia, Blood, Erosion, Lymphoid Hyperplasia Vs Normal
Ours [D2]	1812	97.63	100	98.8	99.2	Inflammatory, Vascular Lesion Vs Normal

As the pathological cues obtained from the input channels (wavelet residuals and colour channels) can exhibit dissimilar features compared to healthy images, one can deduce that the encoder-decoder network shall result in an image close to the input image for healthy classes in each of the six channels. Thus, the difference between the input and the output image with six channels for the normal image set is expected to ideally result in zero difference if reconstructed pixel-wise. Assuming zero difference for healthy images when reconstructed, we incorporate a pixel-wise loss to boost learning of normal image classes through explicit supervision. Given a six-channel input image I, the final residual representation obtained from the encoder-decoder block results in a low pixel-wise reconstruction loss L_r, which can be represented as:

$$L_r = \frac{1}{h\ ch} \sum_{I_i \in h_b} \sum_{ch=1}^{6} ||C_i||_1, \tag{3}$$

for h healthy images in a batch of size b, ch being the total number of input channels (i.e., $ch = 6$) and C_i is the channel wise difference between the input and output for image i.

The output of the final decoder block using a *Tanh* activation layer provides an activation map corresponding to the residual cues that are different in pathological images against healthy images. We, therefore, employ it as an attention mechanism to guide the auxiliary classifier to learn the final classification layer for our proposed model. Figure 4 presents a sample illustration of the cues obtained from the proposed approach, and it can be noted that the visual cues learnt by the network localize the regions of pathology, helping the network in better classification. Finally, we employ an auxiliary classifier to enforce supervision using the input labels in each batch. We simply employ Negative Log-Likelihood Loss as an auxiliary loss function L_a to measure the error between the true labels versus the predicted labels. Thus, our proposed approach makes use of three different loss functions in the training stage as given by Equation 4.

$$L = \lambda_1 L_r + \lambda_2 \sum_{k \in \{E5-D4\}} L_t^k + \lambda_3 L_a, \tag{4}$$

where λs are regularization parameters set to $\lambda_1 = \lambda_2 = \lambda_3 = 1$.

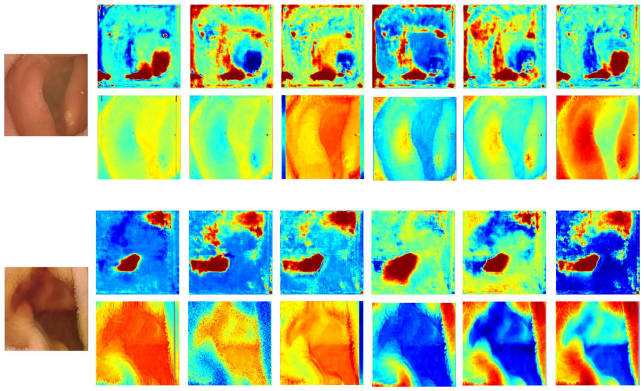

FIGURE 4. Visual multi-channel attention cues obtained for an example image for Lymphoid Hyperplasia (top) and bleeding (bottom) through proposed network. The left image represents the input image, the top row represents the cues learnt for each input channel and the bottom row represents the corresponding inputs in multi-channel input. *Note-The images are shown in colour-map for better visualization.

III. DATASETS

We employ two databases, Kvasir-Capsule dataset (D1) [1] and the Computer Assisted Diagnosis for Capsule Endoscopy Database (CAD-CAP) from the Giana Endoscopic Vision Challenge (D2) [22] to demonstrate the applicability of the proposed approach. We consider nine different classes with different pathologies such as erythema, blood, erosion, angiectasia, ulcer, Lymphangiectasia, polyp, in addition to 2000 randomly sampled clean images from the normal class. (The classes reduced mucosal view (GI debris) and foreign body (capsule endoscope from previous examination) have been excluded from our experiments since they represent neither pathology nor normal scenario). While the Kvasir-Capsule dataset [1] is heavily imbalanced, Giana Endoscopic Vision Challenge [22] provides balanced samples for a set of two pathological conditions and one normal class. Specifically, D2 provides 607 inflammatory images, 605 vascular lesion images and 600 normal images organized in three classes. Irrespective of the dataset and samples for each pathology, we employ a train-validation-test split of 40:10:50 across all experiments.

IV. EXPERIMENTS AND RESULTS

A. TRAINING DETAILS

All the training and testing was conducted on an Nvidia 2080 Ti GPU enabled computer with Linux operating

TABLE 2. Per-pathology sensitivity against the normal class for each of the nine classes in datasets D1 and D2.

Abnormality	Sensitivity [%]	Specificity [%]	AUC [%]	Accuracy [%]
D1				
Angiectasia	97.8	99.6	98.7	99.1
Blood	100	100	100	100
Erosion	94.9	99.3	97.1	98.5
Erythematous	97.4	100	98.7	99.7
Lymphoid Hyperplasia	98.9	100	99.5	99.7
Polyp	96.8	100	98.4	99.9
Ulcer	98.6	100	99.3	99.6
D2				
Inflammatory	99.6	97.3	98.4	98.5
Vascular Lesion	99.3	98.3	98.8	98.8

TABLE 3. Pathology-wise performance comparison for pathology sensitivity against normal, in comparison with other recent works with same or similar class of pathology in WCE (metrics below are as reported in original works).

Approach	Accuracy	Sensitivity	Specificity
Bleeding			
Jia et. al (2016) [20]	99.9	99.2	–
Pan et. al (2011) [18]	–	93.1	96
Aoki et. al (2020) [28]	99.89	96.63	99.96
Aoki et. al (2021) [29]	–	**100**	**–**
Afonso et. al (2021) [30]	–	98.3	98.4
Ours	**100**	**100**	**100**
Erosions and Ulcerations			
Fan et. al (Ulcer, 2018) [31]	95.16	96.8	94.79
Wang et. al (2019) [32]	90.1	89.71	90.48
Aoki et. al (2021) [29]	–	**100**	**–**
Fan et. al (Erosion, 2018) [31]	95.34	93.67	95.98
Aoki et. al (2019) [33]	90.8	88.2	90.9
Ours (Ulcer)	**99.59**	**98.56**	**100**
Ours (Erosion)	**99.32**	**94.88**	**98.56**
Polyp (or Protruding Lesions)			
Yuan et. al (2017) [34]	98	–	–
Aoki et. al (2021) [29]	–	**99**	**–**
Ours	**99.91**	**96.78**	**100**
Angiectasia			
Noya et. al (2017) [35]		89.51	96.8
Leenhardt et. al (2018) [12]		**100**	96
Tsuboi et. al (2019) [36]	–	98.8	98.4
Aoki et. al (2021) [29]	–	97	–
Ours	**99.08**	**97.88**	**99.55**

system (Ubuntu 20.04). The training was conducted with 30 epochs, with a batch size of 32 and a base learning rate of 5e-4. Three milestone epochs were used at epoch number 5, 8 and 12 to decrease the learning rate by 0.3 to converge the learning faster using multi step learning rate scheduler from PyTorch.

TABLE 4. Multiclass pathology classification on Giana Endoscopic Vision Challenge (D2) [22].

Abnormality	Sensitivity	Specificity	AUC	Accuracy
Valerio et.al (2019) [11]	–	–	87	**93**
Vats et.al (2021) [17]	56	59	–	58.7
Ours	**91.38**	**95.52**	**93.52**	91.35

B. ABNORMALITY DETECTION RESULTS

We first perform abnormality detection as a binary classification problem where all pathologies constitute a single class of abnormality. The other class is of normal images. The results can be be seen in Table 1. We compare our approach with Iakovidis *et al.* [27] and Ding *et al.* [8] who have also performed similar binary classification in WCE. While a direct comparison across our work and earlier works [27], [8] is not possible due to the absence of a common dataset, our approach significantly closes the gap for abnormality sensitivity as well as achieves high specificity for normal cases. In comparison to Iakovidis *et al.* [27] with similar scale of data (Table 1. 40% of 1812 and 5489), and Ding *et al.* [8] with much more data (approx 100 times more images), we perform better than both the earlier approaches under limited training data (40% train and 10% validation), indicating robust representations for filtering suspected abnormality.

C. PER-PATHOLOGY RESULTS

Different pathologies exhibit different levels of complexities, due to varying visual appearances and scales as compared to normal images, for example blood pathology is easier to distinguish due to large changes in overall color in the image whereas small scale inflammation or lesions may be harder. Thus resulting classification models may have varying sensitivities for different pathologies. We test if the model generalization extends to different pathologies through multiple binary classifications against each of the commonly known pathology classes in WCE. Table 2 shows the results. The sensitivity in all cases (except erosion) equals or exceeds 97%. Furthermore, the high specificity of our

approach on all classes (especially the fine-grained classes like Erosion and Erythema) can be attributed to the residuals being strong indicators of pathological structures, despite the presence of possible confounding factors like debris and bubbles. This can also be observed from Figure 4 which localizes the residuals helping in classifying the pathology better.

Next, Table 3. presents a comparison of pathology-wise sensitivities with other recent works in WCE. As seen, our work surpasses on sensitivity baseline from most other works and shows competitive performance to Aoki *et al.* [29] who use a dataset of over 66,000 images (approximately 12 times larger than ours).

D. MULTICLASS PATHOLOGY CLASSIFICATION

One crucial aspect for computer-assisted-diagnosis under WCE is when multiple challenging pathologies can be reliably discriminated within a single end-to-end network. To evaluate this, we perform a multi-class pathology classification to evaluate the strength of features in discriminating such categories as the inflammatory and vascular lesion [17] (in contrast to only discriminating pathology from normal). We provide a comparison of our results with other recent works on the same dataset D2 [11], [17] as presented in Table 4. It can be noted that our method surpasses Vats *et al.* [17] by a large margin and closes the gap with Valerio *et al.* [11], who also perform the same classification with features transferred from ImageNet [23], but with a larger (DenseNet161) architecture. The superiority of the results can be attributed to residual features and colour channels in a multi-channel ResNet architecture helping learn discriminative features.

V. EXPLANABILITY ANALYSIS

A commonly employed way of gaining trust in model predictions is to examine its behavior through visual explanations and checking if this agrees with domain knowledge/intuition about the problem. We test the behavior of our classifier in the locality of test samples using 'Local Interpretable Model-agnostic Explanations' (LIME) [37]. LIME works by using the classifier to be evaluated as an input and generates post-hoc explanations contributing to a prediction. For images, this amounts to visualizing image components contributing positively to the true class. We visualize these for a random selection of abnormal images from D2 for the multi-class pathology classifier (Table 4) in Fig.5. The image with abnormality show in yellow circle can be seen in the first columns (Image), followed by components that weight positively to pathology class (middle column). The visualizations show that image regions with abnormality (yellow circle) factor in positively to pathology, whereas seemingly normal regions contribute to normalcy. Interestingly also, we see that visually pervasive pathologies (row 4. and 5.) can be explained with sparser components within the image, whereas obscure cases like in row 2. and 3. includes comparatively denser explanatory components.

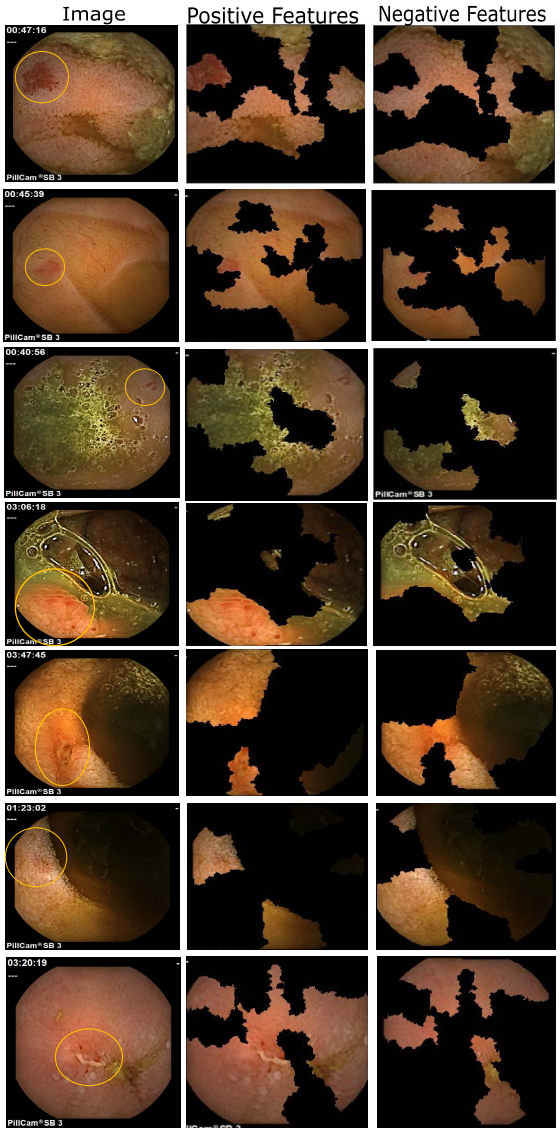

FIGURE 5. Figure shows image components with positive and negative contribution to the ground truth pathology class. The pathology in left column is indicated within the circle. The middle column shows superpixels with positively affect prediction towards true class, the right column shows superpixels that contribute to non-pathology/normal class.

We also visualized components contributing negatively to abnormality (positively to the normal class), these can be seen in the right column. In line with expectation, most of the right column's components comprise of non-pathology subregions within images.

VI. ABLATION: CHANNELS

In this section, we study the impact of dropping different channels on the model's performance. Table 5 shows the classification performance by dropping one of the 6 channels. Similarly, Fig. 6 shows the change in the relative contribution of superpixels to the true class as each of the channels are sequentially zeroed out.

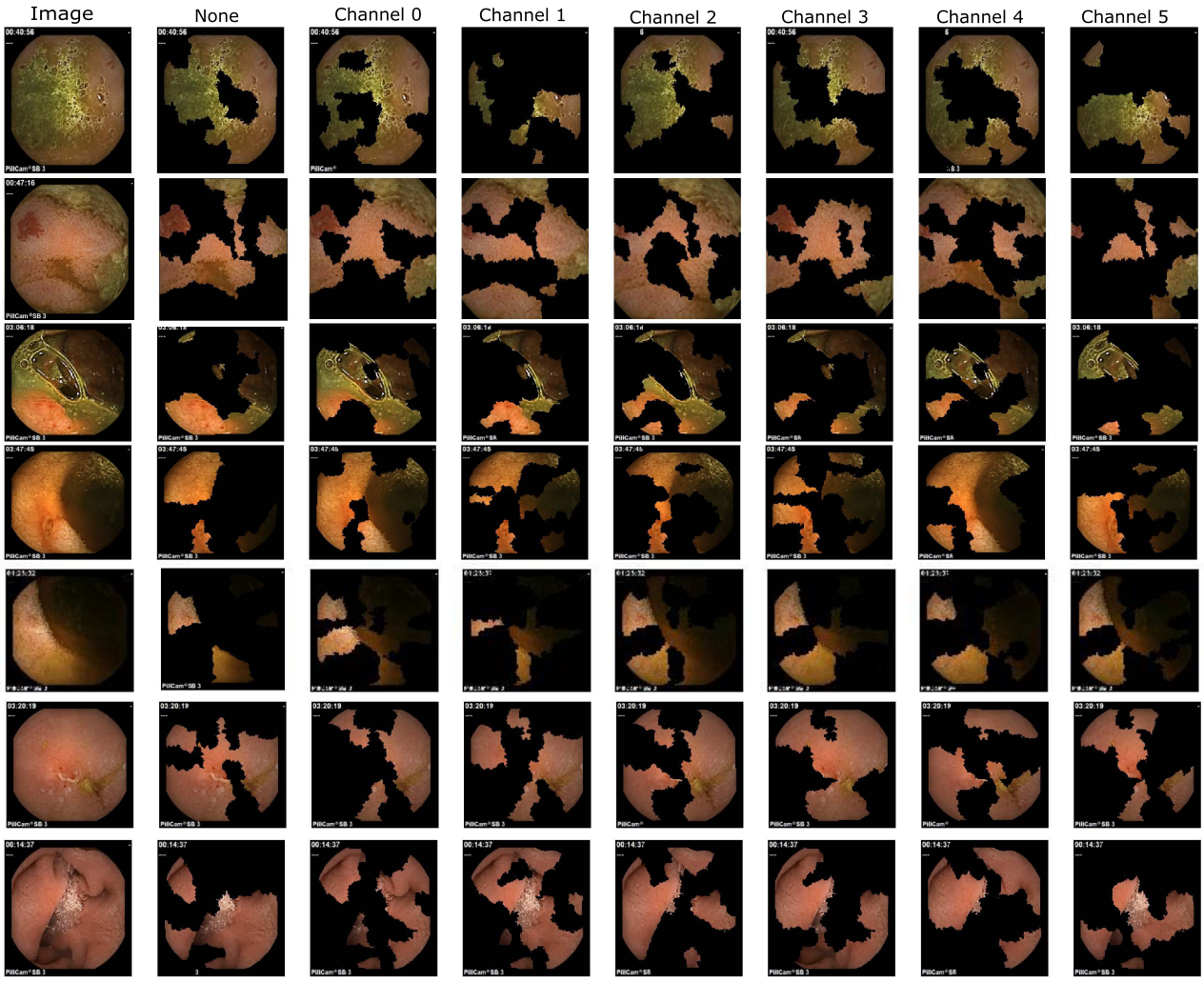

FIGURE 6. Illustration of effect of dropping different channels on model explanations. Column headings indicate the dropped channel. We observed that dropping channel 5 produces consistently sparser explanation as channel 5 encodes low level frequency details of input, while channel 0 exhibits higher pathology sensitivity.

TABLE 5. Impact of each channel on model performance. Evaluated for Multi-pathology classification on D2.

Channel 0	Channel 1	Channel 2	Channel 3	Channel 4	Channel 5	Accuracy	Sensitivity	AUC
x	✓	✓	✓	✓	✓	36.9	52.8	37.1
✓	x	✓	✓	✓	✓	28.3	46.3	28.5
✓	✓	x	✓	✓	✓	33.5	50	33.3
✓	✓	✓	x	✓	✓	72.8	72.9	79.6
✓	✓	✓	✓	x	✓	40.0	39.8	54.9
✓	✓	✓	✓	✓	x	36.6	36.8	52.6
✓	✓	✓	✓	✓	✓	**91.35**	**91.38**	**93.52**

VII. ABLATION STUDY : IMPACT OF LOSS FUNCTIONS

We study the impact of the different loss functions in the proposed approach in this section. Table 6 shows performance obtained for each loss function and the combination of losses used in this work. For the sake of simplicity, we present the ablation studies on dataset D2 alone, where the extremes of

TABLE 6. Performance comparison on dataset D2 for different combination of the three losses, Multi-channel Pixel-wise reconstruction (L_r), Negative Log Likelihood (NLL) (L_a) and feature-level triplet (L_t) loss.

Loss Type			Normal vs Inflammatory		Normal vs Vascular lesion	
L_a	L_t	L_r	Accuracy	Sensitivity	Accuracy	Sensitivity
✓			0.99	0.98	0.98	0.97
	✓		0.43	0.65	0.5	0
		✓	0.46	0.53	0.5	0
✓		✓	0.98	0.99	0.99	0.99
✓	✓		1	1	0.99	0.99
	✓	✓	0.42	0.69	0.5	0

the loss function behaviour can be observed. The Negative Log-Likelihood (NLL) loss (L_a) which operates solely on the embedding of the ultimate layer, is the dominant discriminatory signal for classification between normal and abnormal classes. However, the sensitivity for hard cases is improved even further with the addition of both feature-level (triplet) L_t and pixel-level (Pix-Reg) L_r discrimination. We further observe that in the absence of the dominant discrimination from the supervised categorical classification loss, neither of the two losses or their combination is particularly useful in classification. The combination (L_a and L_t) tends to almost perfect performance, therefore we interpret the results more strictly in this case, as too good a performance may indicate a lack of generalizability or be caused by overfitting (for e.g., to samples coming from the same patient exhibiting similarities). We suspect the learning to benefit additionally from the regularization effect coming from L_r as seen in the combination (L_a, L_r), hence the loss being a weighted average of all three $L_a . L_r, L_t$.

VIII. CONCLUSION

We have proposed a multi-channel cascaded encoder-decoder network in this work for learning a generalizable fine-grained pathological classifier. Noting the differences in the colour appearance of pathology in different directions and orientations, we have proposed to extract residual cues by extracting the differences in wavelets across horizontal, vertical, and diagonal directions in addition to using regular colour channels. The proposed encoder-decoder network fully exploits the multi-channel inputs for learning a generalizable pathology classifier. By employing 40% data (fewer than 2500 labels) for training, we have demonstrated the generalizability of the proposed approach on two WCE datasets (two different capsule modalities) and provided a comprehensive benchmark for different pathologies within a single framework. Our results are better than the state-of-the-art in sensitivity for abnormality detection and fine-grained classification on several of nine different pathologies with much fewer labels and establish a new benchmark with a specificity higher than 97% across all pathology classes.

REFERENCES

[1] P. H. Smedsrud *et al.*, "Kvasir-capsule, a video capsule endoscopy dataset," *Sci. Data*, vol. 8, May 2020, Art. no. 142.
[2] M. L. Ganz, R. Sugarman, R. Wang, B. B. Hansen, and J. Håkan-Bloch, "The economic and health-related impact of Crohn's disease in the United States: Evidence from a nationally representative survey," *Inflammatory Bowel Diseases*, vol. 22, no. 5, pp. 1032–1041, May 2016.
[3] R. Rankala, K. Mattila, M. Voutilainen, and A. Mustonen, "Inflammatory bowel disease-related economic costs due to presenteeism and absenteeism," *Scand. J. Gastroenterol.*, vol. 56, no. 6, pp. 687–692, 2021.
[4] F. Mehta, "Report: Economic implications of inflammatory bowel disease and its management," *The Amer. J. Managed Care*, vol. 22, no. 3, pp. 51–60, 2016.
[5] D. A. Corley and R. M. Peek, "When should guidelines change? A clarion call for evidence regarding the benefits and risks of screening for colorectal cancer at earlier ages," *Gastroenterology*, vol. 155, no. 4, pp. 947–949, Oct. 2018.
[6] G. C. Nguyen, C. A. Chong, and R. Y. Chong, "National estimates of the burden of inflammatory bowel disease among racial and ethnic groups in the United States," *J. Crohn's Colitis*, vol. 8, no. 4, pp. 288–295, Apr. 2014.
[7] J. M. Dahlhamer, E. P. Zammitti, B. W. Ward, A. G. Wheaton, and J. B. Croft, "Prevalence of inflammatory bowel disease among adults aged ≥18 years—United States, 2015," *Morbidity Mortality Weekly Rep.*, vol. 65, no. 42, pp. 1166–1169, Oct. 2016.
[8] Z. Ding, H. Shi, H. Zhang, L. Meng, M. Fan, C. Han, K. Zhang, F. Ming, X. Xie, H. Liu, and J. Liu, "Gastroenterologist-level identification of small-bowel diseases and normal variants by capsule endoscopy using a deep-learning model," *Gastroenterology*, vol. 157, no. 4, pp. 1044–1054, 2019.
[9] 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 Rev. Gastroenterol. Hepatol.*, vol. 9, no. 2, pp. 217–235, 2015.
[10] R. Eliakim, "Video capsule colonoscopy: Where will we be in 2015?" *Gastroenterology*, vol. 139, no. 5, pp. 1468–1471, 2010.
[11] M. T. Valério, S. Gomes, M. Salgado, H. P. Oliveira, and A. Cunha, "Lesions multiclass classification in endoscopic capsule frames," *Proc. Comput. Sci.*, vol. 164, pp. 637–645, Jan. 2019.
[12] R. Leenhardt, P. Vasseur, C. Li, J. C. Saurin, G. Rahmi, F. Cholet, A. Becq, P. Marteau, A. Histace, and X. Dray, "A neural network algorithm for detection of GI angiectasia during small-bowel capsule endoscopy," *Gastrointestinal Endoscopy*, vol. 89, no. 1, pp. 189–194, Jan. 2019.
[13] A. Mohammed, S. Yildirim, I. Farup, M. Pedersen, and Ø. Hovde, "Y-Net: A deep convolutional neural network for polyp detection," 2018, *arXiv:1806.01907*.
[14] J. Yanase and E. Triantaphyllou, "The seven key challenges for the future of computer-aided diagnosis in medicine," *Int. J. Med. Informat.*, vol. 129, pp. 413–422, Sep. 2019.
[15] S. Soffer, E. Klang, O. Shimon, N. Nachmias, R. Eliakim, S. Ben-Horin, U. Kopylov, and Y. Barash, "Deep learning for wireless capsule endoscopy: A systematic review and meta-analysis," *Gastrointestinal Endoscopy*, vol. 92, no. 4, pp. 831–839, 2020.
[16] V. Prasath, "Polyp detection and segmentation from video capsule endoscopy: A review," *J. Imag.*, vol. 3, no. 1, p. 1, 2017.
[17] A. Vats, M. Pedersen, A. Mohammed, and Ø. Hovde, "Learning more for free—A multi task learning approach for improved pathology classification in capsule endoscopy," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2021, pp. 3–13.
[18] G. Pan, G. Yan, X. Qiu, and J. Cui, "Bleeding detection in wireless capsule endoscopy based on probabilistic neural network," *J. Med. Syst.*, vol. 35, no. 6, pp. 1477–1484, Dec. 2011.
[19] H. Saito, T. Aoki, K. Aoyama, Y. Kato, A. Tsuboi, A. Yamada, M. Fujishiro, S. Oka, S. Ishihara, T. Matsuda, M. Nakahori, and S. Tanaka, "Automatic detection and classification of protruding lesions in wireless capsule endoscopy images based on a deep convolutional neural network," *Gastrointestinal Endoscopy*, vol. 92, no. 1, pp. 144–151, 2020.
[20] X. Jia and M. Q.-H. Meng, "A deep convolutional neural network for bleeding detection in wireless capsule endoscopy images," in *Proc. 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2016, pp. 639–642.
[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[22] R. Leenhardt *et al.*, "CAD-CAP: A 25,000-image database serving the development of artificial intelligence for capsule endoscopy," *Endoscopy Int. Open*, vol. 8, no. 3, pp. E415–E420, Mar. 2020.
[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015* (Lecture Notes in Computer Science), vol. 9351, N. Navab, J. Hornegger, W. Wells, and A. Frangi, Eds. Cham, Switzerland: Springer, 2015, doi: 10.1007/978-3-319-24574-4_28.
[25] H. Feng, Z. Hong, H. Yue, Y. Chen, K. Wang, J. Han, J. Liu, and E. Ding, "Learning generalized spoof cues for face anti-spoofing," 2020, *arXiv:2005.03922*.
[26] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.
[27] D. K. Iakovidis, S. V. Georgakopoulos, M. Vasilakakis, A. Koulaouzidis, and V. P. Plagianakos, "Detecting and locating gastrointestinal anomalies using deep learning and iterative cluster unification," *IEEE Trans. Med. Imag.*, vol. 37, no. 10, pp. 2196–2210, Oct. 2018.
[28] T. Aoki, A. Yamada, Y. Kato, H. Saito, A. Tsuboi, A. Nakada, R. Niikura, M. Fujishiro, S. Oka, S. Ishihara, T. Matsuda, M. Nakahori, S. Tanaka, K. Koike, and T. Tada, "Automatic detection of blood content in capsule endoscopy images based on a deep convolutional neural network," *J. Gastroenterol. Hepatol.*, vol. 35, no. 7, pp. 1196–1200, Jul. 2020.
[29] T. Aoki *et al.*, "Automatic detection of various abnormalities in capsule endoscopy videos by a deep learning-based system: A multicenter study," *Gastrointestinal Endoscopy*, vol. 93, no. 1, pp. 165–173, 2021.
[30] J. Afonso, M. M. Saraiva, J. P. S. Ferreira, T. Ribeiro, H. Cardoso, and G. Macedo, "Performance of a convolutional neural network for automatic detection of blood and hematic residues in small bowel lumen," *Digestive Liver Disease*, vol. 53, no. 5, pp. 654–657, May 2021.
[31] S. Fan, L. Xu, Y. Fan, K. Wei, and L. Li, "Computer-aided detection of small intestinal ulcer and erosion in wireless capsule endoscopy images," *Phys. Med. Biol.*, vol. 63, no. 16, Aug. 2018, Art. no. 165001.
[32] S. Wang, Y. Xing, L. Zhang, H. Gao, and H. Zhang, "A systematic evaluation and optimization of automatic detection of ulcers in wireless capsule endoscopy on a large dataset using deep convolutional neural networks," *Phys. Med. Biol.*, vol. 64, no. 23, Dec. 2019, Art. no. 235014.
[33] T. Aoki, A. Yamada, K. A. M. Math, H. Saito, A. Tsuboi, A. Nakada, R. Niikura, M. Fujishiro, S. Oka, S. Ishihara, T. Matsuda, S. Tanaka, K. Koike, and T. Tada, "Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network," *Gastrointestinal Endoscopy*, vol. 89, no. 2, pp. 357–363, 2019.
[34] Y. Yuan and M. Q.-H. Meng, "Deep learning for polyp recognition in wireless capsule endoscopy images," *Med. Phys.*, vol. 44, no. 4, pp. 1379–1389, 2017.
[35] F. Noya, M. A. Álvarez-González, and R. Benitez, "Automated angiodysplasia detection from wireless capsule endoscopy," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2017, pp. 3158–3161.
[36] A. Tsuboi, S. Oka, K. Aoyama, H. Saito, T. Aoki, A. Yamada, T. Matsuda, M. Fujishiro, S. Ishihara, M. Nakahori, K. Koike, S. Tanaka, and T. Tada, "Artificial intelligence using a convolutional neural network for automatic detection of small-bowel angioectasia in capsule endoscopy images," *Digestive Endoscopy*, vol. 32, no. 3, pp. 382–390, Mar. 2020.
[37] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1135–1144.

ANUJA VATS received the master's degree in automotive electronics from Amrita University, India, in 2018. She is currently pursuing the Ph.D. degree with the Department of Computer Science, Norwegian University of Science and Technology (NTNU), Norway. Her research interests include machine learning and computer vision, with emphasis on medical imaging.

KIRAN RAJA received the Ph.D. degree in computer science from the Norwegian University of Science and Technology (NTNU), Norway, in 2016. He is currently a Faculty Member of the Department of Computer Science, NTNU. He was/is participating in EU projects SOTAMD, iMARS, and other national projects. He also works as a Consultant for various national agencies within Norway. He has authored several articles in his field of interest. His main research interests include statistical pattern recognition, image processing, and machine learning with applications to biometrics, security, and privacy protection. He is a member of EAB and the Chair of the Academic Special Interest Group at EAB. He serves as a reviewer for number of journals and conferences.

MARIUS PEDERSEN (Member, IEEE) received the bachelor's degree in data engineering and the master's degree in media technology engineering from the Gjøvik University College, in 2006 and 2007, respectively, and the Ph.D. degree in color image technology from the University of Oslo, in 2011. He is currently a Professor at the Department of Computer Technology and Informatics, NTNU, Gjøvik. He is also the Head of The Norwegian Laboratory for Color and Visual Processing (the Color Laboratory). His work is aimed at quality assessment of images, especially using algorithms.

AHMED MOHAMMED (Member, IEEE) received the master's degree in electronics and information engineering from Chonbuk National University, South Korea, in 2014, and the Ph.D. degree in computer science from the Norwegian University of Science and Technology (NTNU), Norway, in 2020. He is currently a Research Scientist at SINTEF Digital and an Adjunct Associate Professor with NTNU. His research interests include machine learning and computer vision, with emphasis on medical imaging and 3D vision for explainable and data-efficient learning.