Journal Pre-proof

Polyp detection on video colonoscopy using a hybrid 2D/3D CNN

Juana González-Bueno Puyal, Patrick Brandao, Omer F. Ahmad, Kanwal K. Bhatia, Daniel Toth, Rawen Kader, Laurence Lovat, Peter Mountney, Danail Stoyanov

PII: S1361-8415(22)00253-5
DOI: https://doi.org/10.1016/j.media.2022.102625
Reference: MEDIMA 102625

To appear in: *Medical Image Analysis*

Received date : 17 June 2021
Revised date : 22 August 2022
Accepted date : 10 September 2022

Please cite this article as: J.G.-B. Puyal, P. Brandao, O.F. Ahmad et al., Polyp detection on video colonoscopy using a hybrid 2D/3D CNN. *Medical Image Analysis* (2022), doi: https://doi.org/10.1016/j.media.2022.102625.

Medical Image Analysis (2022)

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Polyp detection on video colonoscopy using a hybrid 2D/3D CNN

Juana González-Bueno Puyal[a,b,*], Patrick Brandao[b], Omer F. Ahmad[a], Kanwal K. Bhatia[b], Daniel Toth[b], Rawen Kader[a], Laurence Lovat[a], Peter Mountney[b], Danail Stoyanov[a]

[a] *Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS), University College London, London W1W 7TY, UK*
[b] *Odin Vision, London W1W 7TY, UK*

ARTICLE INFO

Article history:
Received 1 May 2013
Received in final form 10 May 2013
Accepted 13 May 2013
Available online 15 May 2013

Communicated by S. Sarkar

Colonoscopy, Polyp Segmentation, Computer Aided Diagnosis, Temporal Segmentation

ABSTRACT

Colonoscopy is the gold standard for early diagnosis and pre-emptive treatment of colorectal cancer by detecting and removing colonic polyps. Deep learning approaches to polyp detection have shown potential for enhancing polyp detection rates. However, the majority of these systems are developed and evaluated on static images from colonoscopies, whilst in clinical practice the treatment is performed on a real-time video feed. Non-curated video data remains a challenge, as it contains low-quality frames when compared to still, selected images often obtained from diagnostic records. Nevertheless, it also embeds temporal information that can be exploited to increase predictions stability. A hybrid 2D/3D convolutional neural network architecture for polyp segmentation is presented in this paper. The network is used to improve polyp detection by encompassing spatial and temporal correlation of the predictions while preserving real-time detections. Extensive experiments show that the hybrid method outperforms a 2D baseline. The proposed architecture is validated on videos from 46 patients and on the publicly available SUN polyp database. A higher performance and increased generalisability indicate that real-world clinical implementations of automated polyp detection can benefit from the hybrid algorithm and the inclusion of temporal information.

1. Introduction

Colorectal cancer (CRC) is one of the most common types of cancer worldwide, accounting for 10% of all forms of cancer (Bray et al., 2018). Early detection, diagnosis and treatment can effectively reduce CRC incidence and mortality (Van Rijn et al., 2006). Colonoscopy is the gold standard screening procedure for early CRC detection, during which the bowel is visually inspected for polyps and cancer using an endoscope (Rex et al., 2009). The risk of interval cancer was found to decrease by 3.0% with each 1.0% increase in the adenoma detection rate (Corley et al., 2014). Unfortunately, colonoscopy outcomes are highly variable, with high inter-operator disparities (Corley et al., 2014) as well as intra-operator dependency, where detection rates decline with fatigue (Leufkens et al., 2012).

Computer-aided detection (CAD) systems to assist endoscopists in polyp detection tasks have been widely researched in the last 30 years. Recently, great progress has been reported (Hassan et al., 2019; Wang et al., 2018a,b) particularly when using approaches based on convolutional neural networks (CNNs) (Brandao et al., 2018, 2017; Tajbakhsh et al., 2015; Wang et al., 2018b). These approaches have reported robust and promising results (Ahmad et al., 2019a) and multiple clinical studies and randomised control trials have begun to evaluate CAD technology as well as to consider ethical and regulatory aspects (Su et al., 2020; Ahmad et al., 2019b). Notably, commercial CAD systems have recently emerged with technology based on deep learning such as GI Genius (Medtronic,USA) (Repici

*Corresponding author
e-mail: j.puyal@ucl.ac.uk (Juana González-Bueno Puyal)

et al., 2020), CADDIE (Odin Vision, UK) (Odin Vision, 2020), CAD EYE (FujiFilm, Japan) (Weigt et al., 2020), DISCOVERY (Pentax Medical, Japan) (Medical, 2019), AI4G (AI4GI Corp., Canada) and (Olympus America, USA) (Chahal and Byrne, 2020), or EndoBrain (Cybernet Systems, Japan) (Kudo et al., 2020). However, one of the main challenges when developing deep learning models is the limited availability of labelled data because full length colonoscopy videos are not usually recorded clinically and when recorded they are logistically challenging to handle (Ahmad et al., 2020). On the contrary, still frames from regions of interest within the procedures are routinely stored in clinical reports enabling still image polyp databases (Bernal et al., 2017; Jha et al., 2020). As a result, while most current CAD systems have been developed using still images, they are translated to endoscopy units where real time videos are used to detect polyps. Videos introduce difficulties as they can present poor visibility and variability in polyp appearance that might lead to a lack of temporal coherence in consecutive frames yielding short, false predictions (Bernal et al., 2017). It is therefore of paramount importance to demonstrate performance on videos and address model behaviour and stability in practical conditions.

Endoscopic videos can be exploited to increase the temporal correlation in the predictions by extracting temporal representations. This line of thought has previously been employed for videos by the means of recurrent neural networks (RNN), such as long short-term memory, 3D CNNs, or two-stream models, demonstrating state-of-the-art performance for action recognition tasks (Carreira and Zisserman, 2017). Similarly, the use of temporal information in endoscopic CAD has been studied for different approaches. For instance, (Ma et al., 2020) show that temporal consistency can be used to reliably detect true positive predictions and apply it for automatic data labelling. Further, temporal architectures have been exploited such as in (Eelbode et al., 2019), where the authors implement an RNN on top of a CNN to improve polyp segmentation accuracy. Likewise, dense 3D networks have been explored such as C3D to classify endoscopic frames containing polyps (Itoh et al., 2018; Misawa et al., 2018; Itoh et al., 2019), a 3D fully convolutional network for polyp segmentation (Yu et al., 2016) or a 3D CNN for polyp structures identification (Liu et al., 2020). Despite the benefits of 3D architectures, a major drawback of this type of models is the need of a large number of training samples, while medical data collection remains a challenge. Various strategies have been proposed to include temporal information leveraging access to a limited number of videos. For example, in (Qadir et al., 2019) a false positive reduction stage was appended to a polyp detection model showing an increase in specificity while limiting the loss in sensitivity. Tracking algorithms can be combined with detection CNNs to temporally refine results but the re-initialisation of the tracker can be problematic (Zhang et al., 2018; Poon et al., 2020). Recently an approach fusing two CNN streams, one receiving the input frame, and the other one optical flow information, was reported but can suffer from errors in the optical flow estimation (Zhang et al., 2019).

This paper presents a hybrid 2D/3D architecture for polyp segmentation in colonoscopy videos. This network combines 2D spatial representations with a third temporal dimension in a hybrid manner. The novel architecture presents the benefits of traditional 2D networks, allowing to leverage small, static image datasets by pre-training the 2D backbone without the need of large video datasets. Additionally, it includes the advantages

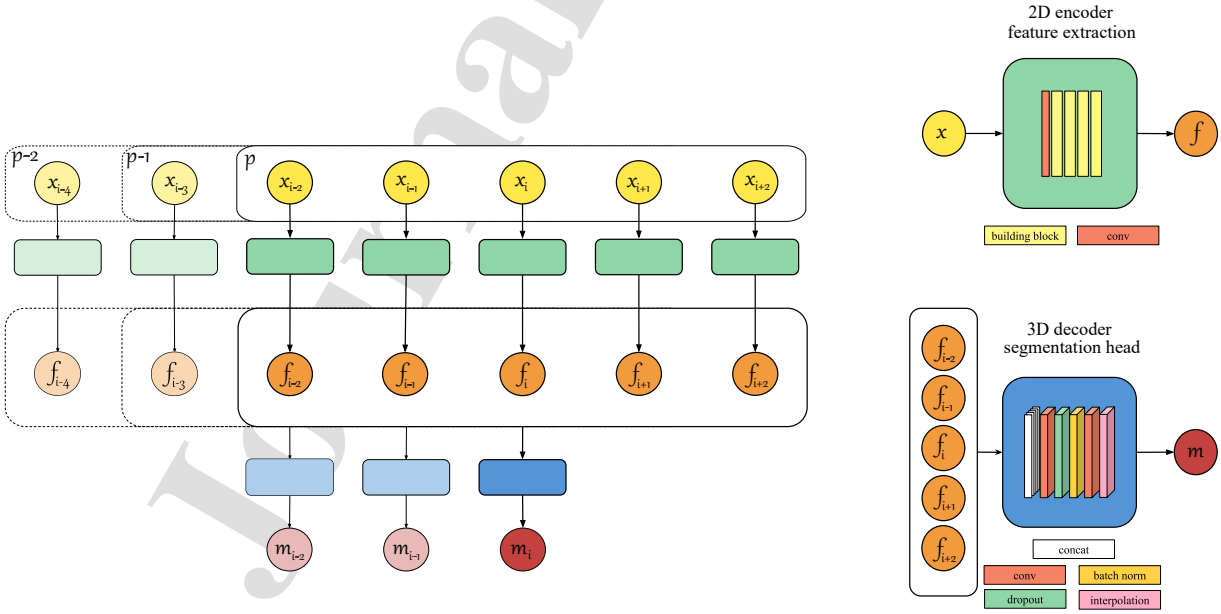


Fig. 1. Architecture of the proposed hybrid segmentation network for a temporal depth or input size of $d = 5$ frames. A set of features f is generated for each input image x with the 2D encoder (depicted in green). d feature sets are given to the 3D decoder (depicted in blue) to generate an output segmentation m.

of a 3D network as it can be trained or fine-tuned on videos in an end-to-end fashion. The use of the hybrid model would yield better representations of the lesions leading to an increase in sensitivity, as well as a decrease of short false positives while introducing bearable delays in the predictions.

Although a segmentation network was employed, the ultimate application was polyp detection. This allows to label polyps with a box around each lesion rather than with delineations, therefore reducing the time and cost of the labelling. Moreover, following a polyp detection event, a polypectomy is performed by resecting the lesion using tools with low maneuverability and precision. As opposed to segmentation metrics, object-wise detection metrics enable a clinically relevant evaluation without focusing on detailed delineations that would nevertheless be difficult to execute during polypectomies. Additionally, from a usability point of view, overlaying a predicted box around a lesion allows a good view of the polyp without occluding it.

The core hybrid model was initially reported at MICCAI and evaluated on internal data (Puyal et al., 2020). In this paper, we expand the analysis of the hybrid architecture and its benefits. Our new contributions include: (i) refinement of the proposed approach in order to determine the optimal temporal window size of the input video; (ii) new evaluation of the proposed method and the baseline on the publicly available SUN polyp database (Misawa et al., 2020) in order to demonstrate generalisation capabilities and provide the basis for comparison with other existing methods; (iii) proposal of novel user-centred metrics to evaluate the performance of the models.

2. Methods

A two-step temporal segmentation algorithm was developed (see Figure 1) (Puyal et al., 2020). The proposed architecture was capable of learning a spatial representation through the 2D encoding stage, allowing to apply transfer learning from larger and more varied 2D datasets. A 3D segmentation stage followed in order to generate temporally coherent polyp segmentation masks.

During training and validation pixelwise predictions and losses were used by converting the ground truth box annotations into binary masks. At inference time, the obtained masks were thresholded and each separate detected region was replaced by a rectangle inscribing the detection, so that the obtained boxes were evaluated on a object-wise level. Figure 2 shows an example of this process.

2.1. Hybrid architecture

The hybrid model contains an initial 2D stage for feature extraction. Any CNN backbone could be utilized, although in our implementation a Resnet-101 architecture was used as the encoder. The Resnet model included a convolutional layer, followed by four sets of building blocks containing 3, 4, 23 and 3 residual blocks, sequentially (He et al., 2016). The last fully-connected layer was removed, the output then consisting of a set of 2048 feature maps per image.

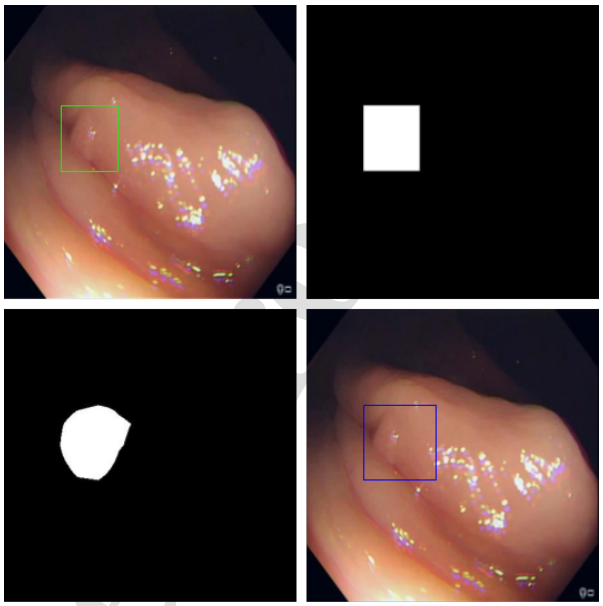


Fig. 2. Example of inputs and outputs to the network. Images are annotated by experts who provide a box ground truth around each polyp (top left). The box is converted into a binary mask for training purposes (top right). On inference the segmentation output for the polyp class is thresholded at 0.5 (bottom left) and each predicted region is converted to a box used for evaluation (bottom right).

The encoder is followed by a 3D segmentation stage to generate a final segmentation output. This 3D decoder is composed of one concatenation layer, two 3D convolutional layers, dropout, batch normalisation and an interpolation layer for upsampling (see Figure 1). This structure is an inflated version of the segmentation head from a Fully Convolutional Network (FCN) (Long et al., 2015), where "inflated" refers 2D layers expanded into 3D (Carreira and Zisserman, 2017). The first convolutional layer uses a kernel of size $[d, 3, 3]$, where d is the temporal depth or number of input images to the network, with a padding and a stride of $[1, 1, 1]$. The second convolutional layer uses the same stride, no padding, and a kernel of $[d - 2, 3, 3]$ and a sigmoid activation layer, generating a segmentation output containing one channel per class with pixel predictions $y \in [0, 1]$. The upsampling layer resizes each channel from the segmentation output m to the size of the input images.

An input sample for the network consists of d consecutive video frames, d being configurable. The output of the model is a segmentation map corresponding to the middle input image (therefore constraining d to be an odd number). The input images are passed through the backbone, extracting a set of spatial features f_i^p for the i_{th} image and p_{th} sample as depicted in Figure 1. The features set for an image have a shape of $[2048, w, h]$, where w and h are the width and height of features, 32 times smaller than the original input images. The first layer of the segmentation head, namely the concatenation step, stacks the features f_i^p for the images $i \in \{1, 2, ...d\}$ from the same sample p into a batch of 3D features. Consequently, the input to the

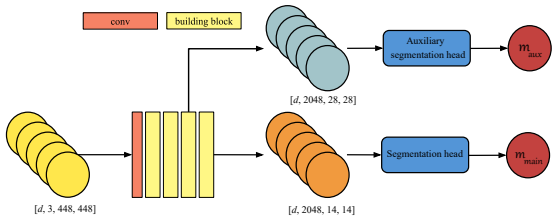


Fig. 3. Training architecture showing the auxiliary segmentation head. Two sets of feature maps are generated per input and inserted in the segmentation heads that generate two outputs m of identical size.

first convolutional layer has a shape of $[P, 2048, d, w, h]$, where P is the number of samples in a batch. After the last layer in the segmentation head, the network generates a probability map m_p per sample, trained to predict the polyp in the image in the middle of the sample.

A multiscale cross-entropy loss was used to train the model, so that the final pixelwise loss was defined as:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{main} + \mathcal{L}_{aux} \\ &= -\sum_{c} c \cdot \log(y_{main}) - \sum_{c} c \cdot \log(y_{aux})\end{aligned} \tag{1}$$

where $\mathcal{L}_{main}$ was the main loss, $\mathcal{L}_{aux}$ was the auxiliary loss, $c = \{0, 1\}$ were the polyp and background ground truth classes, and y_{main} and y_{aux} were each of the pixel predictions from the segmentation outputs m_{main} and m_{aux} arising from the main and auxiliary heads, respectively. The final loss was obtained by averaging the pixelwise losses. As depicted in Figure 3, during training, an additional auxiliary segmentation head received features from the third backbone building block. These feature maps undergo fewer pooling steps, and therefore are twice the size of the main feature maps, as shown in Figure 3 for an image input size of 448x448. After passing through the segmentation head both generated outputs have an equal size due to the interpolation layer. This training strategy was adapted from (Long et al., 2015) where three feature maps were used to combine fine and coarse layers. In this work two feature maps were used instead to make the models less granular and more efficient, and the losses were computed separately and combined, rather than combining the feature maps previous to computing the final loss.

The model was trained and tested on an end-to-end fashion. At train time, the inputs were randomised so an input batch of P random samples contained N images, where $N = d \times P$. At inference time, a prediction was obtained for each video frame by running the model in a sliding window fashion with a stride of one and d frames per sample. Runtime was improved by encoding each image once with the 2D feature extraction and storing the generated features f_i^p for each image to be used by the 3D segmentation head during the following $d - 1$ frames, saving time and computational resources.

2.2. Training strategy

Random sampling of 5000 samples was performed on each epoch to minimise overfitting, re-sampling at every new epoch. Data augmentation was applied in such a way as to guarantee identical augmentations within samples. The augmentation operations consisted of random affine transformations (rotation, translation and scaling) and random colour transformations (brightness, contrast and saturation). Finally, the images were pre-processed by cropping out the video borders, followed by resizing the images to 448 by 448 pixels, and an intensity normalization step.

All available positive images were used during training, and a data mining strategy was adopted for adding beneficial negative images to the training set as seen in (Podlasek et al., 2021). An initial model was trained uniquely with positive samples and was used for inference on the available set of negative images (from training procedures), which was shuffled randomly. Images yielding false positives were selected until reaching 15% of the new training set. This data mining strategy was selected empirically, after comparing it to random and semi-random negative data selection.

Cross-entropy loss was used for the experiments and Adagrad for optimisation. Two output classes were defined: polyp and no-polyp presence. The epoch with the highest pixel accuracy in the validation set was selected for testing. All models were trained with PyTorch on an NVIDIA Tesla V100 DGXS 32GB GPU. The batch size and sample size were adapted depending on the length of the input window d so that the largest possible batch would fit in a single GPU's memory. The inference speed increased linearly with the number of input frames d (the model predicted at 19 frames per second with $d = 5$ on the training GPU).

3. Experimental results

3.1. Datasets

The data was divided into two separate datasets: the *Video Dataset* composed of consecutive video frames and the *Image Dataset* composed of static images.

Video Dataset A series of 95 videos, from 95 patients, which was collected in University College London Hospital with an Olympus EVIS LUCERA endoscope under ethics REC reference 18/EE/0148. A total of 234 histologically confirmed polyps were extracted into single-polyp video sequences. The frames in these sequences were annotated by expert colonoscopists by drawing bounding boxes around each polyp. Only frames showing polyps, captured in white light imaging mode were included. The 25 full-length negative videos were added to the testing set, whereas the 70 procedures containing polyps were randomly split into training, validation and testing sets. The data was split on a per-procedure basis, where each procedure corresponded to a different patient, ensuring there was no patient data overlap between the sets of data. 51,426 frames from 173 polyps within 45 procedures were used for training and 2,152 frames from 8 polyps within 4 procedures were used for validation. 20,943 frames from 21 procedures and 53 polyps were used for testing, as well as 542,583 non-polyp frames from the 25 negative procedures.

SUN dataset Additionally, to test the generalisation capabilities of the model, the SUN Colonoscopy Video Database (Misawa et al., 2020) was used exclusively for testing. This dataset

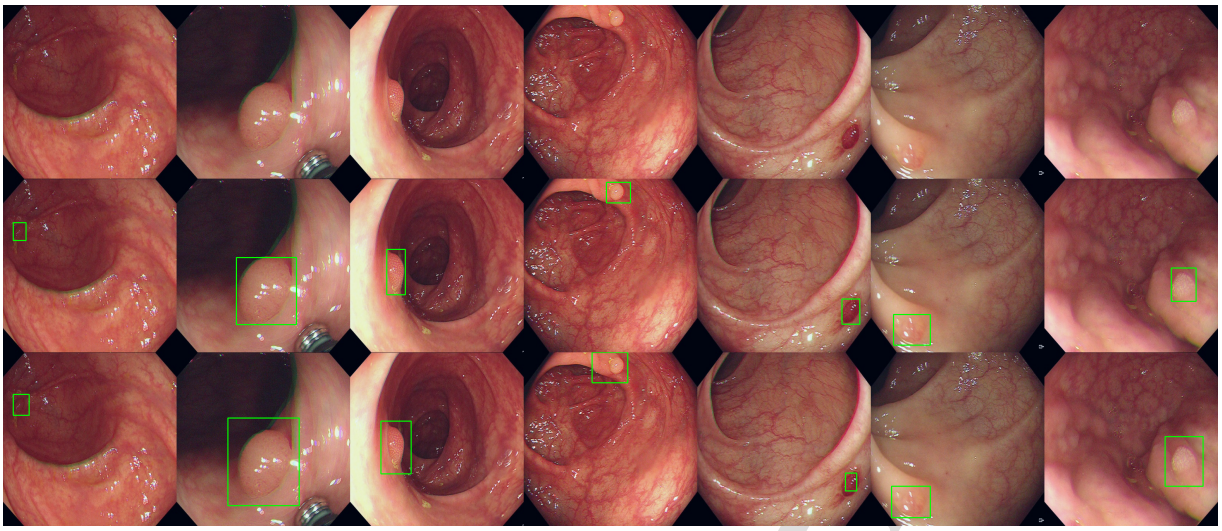


Fig. 4. Examples on SUN database. Input images (top), Ground truth annotations (middle) and Hybrid model predictions (bottom).

Table 1. Evaluation of baseline and proposed methods on the *Video Dataset* (*pp* and *pf* denote per-polyp and per-frame sensitivity, respectively). Object-wise metrics were used for the assessment.

Method	Sens (*pp*)(%)	Sens (*pf*)(%)	Spec (%)	Prec (%)	F1 (%)	Dice (%)	Δ A-corr (%)	TC (%)
FCN (ImageNet)	100.00	83.56	83.04	88.11	85.78	69.68	20.50±16.16	79.55
Hybrid (FCN)	100.00	85.66	83.60	93.27	89.30	**74.08**	**11.92±11.97**	84.24
Hybrid (ImageNet)	100.00	**86.14**	**85.32**	**93.45**	**89.65**	73.48	12.24±11.74	**84.64**

is composed by 49,136 polyp frames from video sequences from 100 polyps, annotated with bounding boxes. It also includes 109,554 frames from non polyp scenes. Further information about the data can be found in (Misawa et al., 2020). Figure 4 shows example polyp images from this dataset.

Image Dataset Static polyp images were gathered from two sources: the publicly available Kvasir dataset (Jha et al., 2020) composed of 1,000 polyp images and corresponding masks, and a dataset containing 833 polyp images collected from reports from University College London Hospital under ethics REC reference 18/EE/0148 and annotated by expert colonoscopists by drawing bounding boxes around polyps. This set of 1,833 white light, polyp images was solely used for training purposes.

3.2. Comparison and evaluation metrics

In order to assess the temporal benefits of the model, its comparable 2D network, an FCN with a Resnet101 backbone, was implemented (Long et al., 2015). Whereas the backbone used was identical to the one in the hybrid model, the segmentation head was a deflated version of the hybrid one. In this case, 3D convolutional, batch normalisation and pooling layers were replaced by their 2D corresponding versions, maintaining all other parameters. During training, an auxiliary segmentation head was used in the same manner as for the hybrid network. The training strategy and parameters for the baseline model were kept identical to the hybrid model, when possible, to ensure comparison fairness.

Object-wise metrics were used for evaluation, namely sensitivity, precision, and F1-score on videos with polyps, and specificity on non-polyp videos. Further implementation details are available in (Bernal et al., 2017). Predicted polyp objects were denoted by a rectangle enclosing pixels classified as polyp at a threshold of 0.5 (see Figure 2). This allowed comparison with the ground truth annotations of rectangular bounding boxes. Dice score was reported on true positive frames to assess the quality of the boxes overlap. Per-polyp sensitivity was also reported, considering a true positive when at least one frame was correctly detected for each polyp.

In order to determine the consistency of the predictions over consecutive frames, temporal coherence (TC) was computed as defined in (Bernal et al., 2017). Additionally, auto-correlation of masks was measured to assess both temporal and spatial correlation between two consecutive mask predictions. The auto-correlation for a given pixel position over a sequence of masks is defined as:

$$r = \frac{\sum_{i=1}^{N-k}(Y_i - \bar{Y})(Y_{i+1} - \bar{Y})}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2} \tag{2}$$

where Y_i is the value of a pixel in a certain position, and $\bar{Y}$ is the average of the pixel values in that position over the entire sequence. After obtaining a 2D vector with auto-correlation values per sequence, the average over the x and y axis was computed. The absolute difference with respect to the ground truth

auto-correlation was computed, and mean and standard deviation were reported.

While assessing object-wise metrics on all test frames is an important performance metric, so is measuring the delay when first detecting a polyp. Shorter delays will have a higher influence on subtle polyps detection that would be otherwise missed. This detection delay was measured with two metrics namely the time to detection and the network reaction time. The network reaction time Δ_N measures the time needed for a polyp to be detected and was defined by the time between its first appearance (t_{app}) and its first detection (t_{det}), averaged over all polyps in the test sequences.

$$\Delta_N = \frac{1}{P}\sum_{P} t_{app} - t_{det} \tag{3}$$

The time to detection Δ_d was defined as the time elapsed between between the first appearance of a polyp and its display to the user, taking into account the additional latency induced during real-time inference, where for an input of 11 frames, it would be necessary to wait to acquire the following 5 frames to make a prediction on the central image.

$$\Delta_d = \frac{1}{P}\sum_{P} t_{app} - t_{det} + \frac{d}{2} + 1 \tag{4}$$

The proposed architecture was additionally evaluated on an external dataset. Results in this dataset were computed according to object-wise metrics, as well as image-wise metrics described in (Misawa et al., 2020) to allow for comparison with other published results. In (Misawa et al., 2020), results were computed per frame, where a prediction for a frame was deemed positive if at least one box was predicted in the frame. This type of image-wise metrics yield higher results than object-wise metrics (described above), as a frame will be deemed as a true positive even when the detection does not overlap with the ground truth polyp. The second difference in the metrics is related to the per-polyp sensitivity, in that, for the external database, a video sequence is deemed positive if at least half of its frames are predicted as positive.

3.3. Results and analysis

3.3.1. 2D / Hybrid comparison

To establish a baseline, a 2D FCN was trained initialising the backbone weights on ImageNet - referred to as *FCN (ImageNet)*. The training set for this model consisted of images from the *Video Dataset* training set and the full *Image Dataset*. Furthermore, 10,000 negative images from the training procedures from the *Video Dataset* were added to the training set using the strategy previously described, by means of an FCN model formerly trained on positive images exclusively. Correspondingly, a hybrid model was trained on the training set from the *Video Dataset* and 10,000 negative images, following the negative mining strategy. The *Image Dataset* was excluded as the hybrid architecture needs consecutive frames as inputs. The hybrid network was trained using the weights from *FCN (ImageNet)* to initialise and freeze the backbone, therefore only training the segmentation head - this experiment was named *Hybrid (FCN)*. An input temporal depth d of five frames was used. Sharing a common backbone, it was possible to solely evaluate the effect of the 3D segmentation head. The proposed model was also trained initialising the backbone from ImageNet weights and training the full network. This experiment was referred to as *Hybrid (ImageNet)*.

Table 1 depicts the associated results when tested on the *Video Dataset* internal testing set. When comparing *FCN (ImageNet)* and *Hybrid (FCN)* it can be observed that the incorporation of the 3D component caused a general increase in performance, particularly in terms of temporal consistency metrics. The temporal coherence increased by 5% showing that predictions were more consistent on consecutive frames. The enhance in temporal correlation was supported by a decrease of 9% in the difference between the auto-correlation for the hybrid predictions and the ground truth masks, indicating that consecutive predicted masks presented a higher similarity. Additionally, the

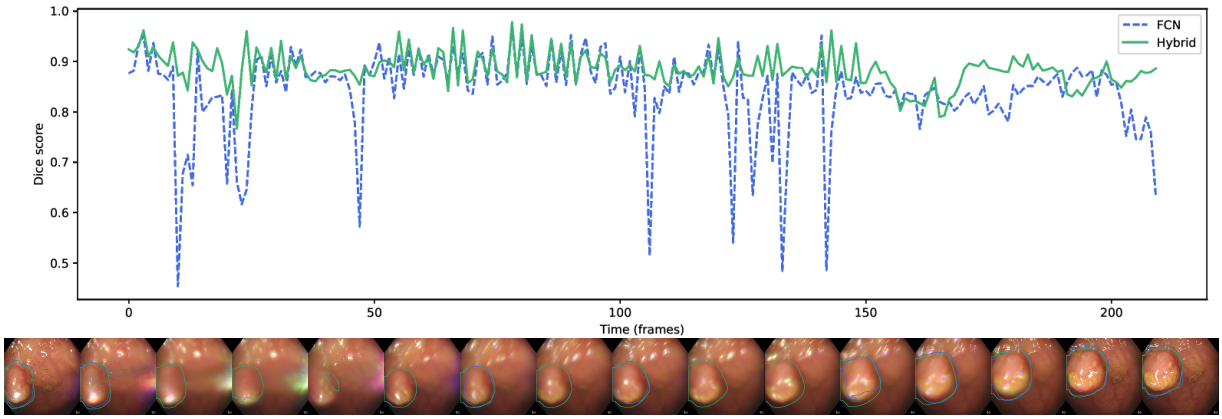


Fig. 5. Results on a polyp sequence showing (top) the box dice overlap with the ground truth and (bottom) segmentation outputs for the FCN (blue) and Hybrid (green) before box conversion postprocessing. The segmentation outputs are shown in blue for the FCN and green for the Hybrid model (for interpretation of the references to colour, the reader is referred to the web version of this article).

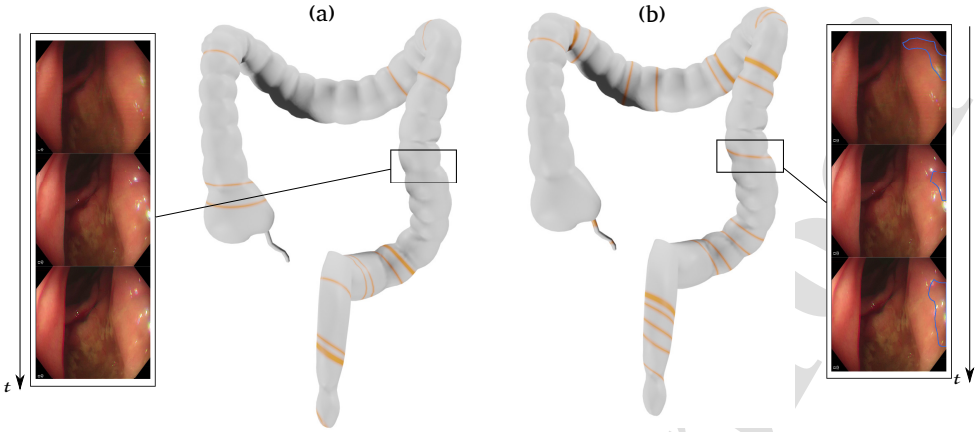


Fig. 6. Prediction timelines for a non-polyp procedure mapped onto a colon model for (a) Hybrid (ImageNet) and (b) FCN (ImageNet), where orange stripes denote false positives. Network outputs are shown as an overlay on a video section.

introduction of temporal components led to an increase of both sensitivity and specificity. This was achieved by a reduction of short false positives and negatives in both hybrid implementations. The gain in sensitivity by the hybrid model came accompanied by a considerable increase of ~ 5% in the dice score for detected polyps, showing that the quality of the segmentation masks benefited from the temporal component. Figure 5(top) shows the dice score over a polyp sequence for the *FCN (ImageNet)* and the *Hybrid (FCN)* models, where it can be observed that the dice score over consecutive polyp frames is more stable for the hybrid model. From the segmentation examples in Figure 5(bottom) it can be noted that both models generated similar outputs on good quality images. However, on blurry frames the FCN yielded false negatives while the hybrid model successfully used information from surrounding frames.

The *Hybrid (ImageNet)* experiment was trained from ImageNet weights, achieving the highest sensitivity and specificity, 86.14% and 85.32% respectively (see Table 1). It also showed the highest F1-score, demonstrating that it was possible to obtain advantages from the hybrid architecture over its 2D counterpart, even when using a lower amount of training data. Our hybrid architecture successfully offered temporal benefits without overfitting, a major disadvantage of 3D models. The results for *Hybrid (ImageNet)* were comparable to *Hybrid (FCN)*, showing similar increases in the quality of the segmentation overlap as well as the temporal consistency metrics. Figure 6 shows the per frame predictions on one of the non-polyp full colonoscopic withdrawals from the testing set, where it can be seen that the number of false positives is reduced throughout the procedure with the hybrid network compared to the FCN. Although the mapping to a 3D model of the colon was not fully realistic, it gave an indication of the clinical importance of reducing the false positives, as too many false alarms can reduce the usability of a clinical system.

A data benchmark was performed to show the closest comparison possible between architectures. Negative mining was used to select non-polyp training frames for experiments reported in Table 1, hence the training images used were not identical between experiments. Table 2 shows the performance of the previous experiments when trained uniquely on positive data, allowing for a better comparison between the FCN and our proposed architecture. Three models were trained as follows: (i) A 2D FCN was trained on the *Video* and *Image* training datasets, initialising its backbone from ImageNet weights. (ii) A hybrid model was trained initialising the backbone from this *FCN (ImageNet)*, and training exclusively the segmentation head. (iii) A hybrid model was also trained from ImageNet, without freezing any layers. It is important to note that these networks were trained exclusively on positive samples and false positives were to be expected. The results presented in Table 2 show that, when pre-training with *FCN (ImageNet)*, the proposed model improved the performance considerably in all aspects when compared to FCN. Particularly, a 20.19% rise in specificity was achieved, along with an important improvement in the auto-correlation of the predictions, showing that the hybrid architecture was able to get rid of false short positives without training on negative images. However, the hybrid model initialised from ImageNet, *Hybrid (ImageNet)*, yielded poorer results when compared to the *FCN ImageNet)*. This could be due to the limited amount of training data used in this instance, the *Hybrid (ImageNet)* also lacking the *Image* dataset from its training set. The results might indicate the tendency of the hybrid architecture to overfit when reducing the dataset excessively, a common problem on 3D architectures. Nevertheless this problem can be solved as the proposed architecture was shown to successfully allowing to pre-train on still images while benefiting from the temporal stability provided by the 3D segmentation head.

3.3.2. External testing

The models were tested on a completely external, publicly available dataset (Misawa et al., 2020) to assess their generalisation capabilities as well as to allow for external comparison, and the results were presented in Table 3. When tested on the SUN dataset our model achieves 86.99% sensitivity and

Table 2. Evaluation of network performance when negative data is not included in the training set. Results are reported on the test set from the *Video Dataset* (*pp* and *pf* denote per-polyp and per-frame sensitivity, respectively). Object-wise metrics were used for the assessment.

Method	Sens (*pp*) (%)	Sens (*pf*) (%)	Spec (%)	Prec (%)	F1 (%)	Dice (%)	Δ A-corr (%)	TC (%)
FCN (ImageNet)	100.00	87.77	54.02	87.81	87.80	72.22	17.12±15.18	84.58
Hybrid (FCN)	100.00	**88.88**	**74.18**	**92.73**	**90.76**	**75.06**	**9.77±9.77**	**87.78**
Hybrid (ImageNet)	100.00	85.79	44.54	87.45	86.61	68.27	10.89±11.25	84.42

Table 3. Evaluation on the SUN polyp database (*pp* and *pf* denote per-polyp and per-frame sensitivity, respectively). † denotes metrics from (Misawa et al., 2020)

Method	pp Sens (%)	pf Sens (%)	Spec (%)	F1 (%)
YOLOv3 (Misawa et al., 2020) †	**98.0**	90.5	**93.7**	-
FCN †	96.0	**91.38**	85.61	87.49
Hybrid (ours) †	93.0	86.99	90.41	**87.58**
FCN	**100.0**	**89.46**	76.50	80.80
Hybrid (ours)	97.0	84.71	**84.57**	**83.29**

90.41% specificity, which is a higher performance than on the *Video Dataset* reported on Table 1 (86% sensitivity and 85% specificity). However, when evaluated with our object-wise metrics (defined in Section 3.2), our model achieves an 84.71% sensitivity and 84.57% specificity, marginally lower than on our testing set. This shows that the hybrid model is able to maintain the test set performance on a completely different dataset and generalises adequately. Nevertheless, the small loss in per-frame sensitivity is translated in a drop of the per-polyp sensitivity, where seven polyps yield false negatives on half or more of their frames. Additionally, the results on the SUN database are compared to other published results on this data. The results from Misawa et al. (Misawa et al., 2020) present a 3% higher sensitivity and specificity when compared to our hybrid model, and a 5% higher per polyp sensitivity. It is important to note that the model from (Misawa et al., 2020) was trained with images from 5 hospitals, one of them being Showa University Northern Yokohama Hospital, where the SUN dataset was acquired. Even though there is no patient overlap, the domain of the SUN dataset is likely to be within the distribution of their training set (due to factors such as the same endoscopic device, similar colonoscopic techniques, colonoscopists overlap, similar patient demographics, etc.). Finally, the FCN baseline model was also tested on the SUN database. It showed a higher sensitivity and a lower specificity when compared to the hybrid model, making it difficult to compare both architectures. However, the F1-score was higher for the hybrid model, showing that this architecture obtained a better sensitivity/specificity balance, possibly due to having learnt a better data representation. When computing the results with our per-lesion metrics, the sensitivity and specificity drop to 89.46% and 76.50% for the FCN model, showing that it has a poorer per-frame precision, yielding false positive detections on polyp frames. All in all, the proposed hybrid model maintains its performance when tested on a dataset with a new distribution, generalising better than its 2D counterpart. When compared to (Misawa et al., 2020) the performance is found to be comparable, demonstrating the generalisation capabilities and robustness of the hybrid architecture.

3.3.3. Input sequence length analysis

Previously presented experiments in Tables 1 and 2 were trained with a temporal depth $d = 5$, meaning that each sample consisted of five consecutive frames. The parameter d was chosen empirically based on the average speed of an endoscope in the colon and the frame rate of the videos used (25fps). Samples of five frames where selected to ensure enough variation of appearance while conserving the same scene within the sample. Following the experiments presented in Section 3.3.1 to ascertain the benefits of the hybrid architecture, further investigation was implemented to find the optimal depth for the 3D stage of the network. Several models were trained with the same parameters as for the *Hybrid (ImageNet)* model from Table 1. Figure 7 shows the performance of the hybrid model when trained with different input sequence length, or temporal depth d, values. The input sequence length ranged from 3 frames, the minimum possible temporal depth, to 41 frames, the maximum possible length allowed by the capacity of the GPU, which corresponds to 1.64 seconds of video feed, enough time for the scene to change radically. A first bar with a temporal depth $d = 1$ was included, corresponding to the *FCN (ImageNet)* baseline.

Figure 7 shows the performance of the different input size experiments in terms of sensitivity, specificity and F1-score, as defined in Section 3.2, as well as for time to detection. As it can be observed the overall performance is consistently higher in experiments where the hybrid architecture is used ($d > 1$), with the exception of $d = 41$. In this last experiment the sensitivity drops to 74%, probably because the input samples mask out the polyp frames. The results become challenging to interpret when comparing the temporal depth for different hybrid experiments. Despite all efforts to maintain reproducibility between experiments, it is important to note that the changes on the input frames modify the order in which data is loaded during training, generating random differences and thus small variations in performance cannot be confidently attributed to the input size d. The sensitivity, specificity and F1-Score metrics are depicted cumulatively in Figure 7 to ease the visualisation. When taking those three metrics into account, the results seem to indicate an increase in performance with longer inputs (between 3 and 25). However, the improvements found were small. For instance, the

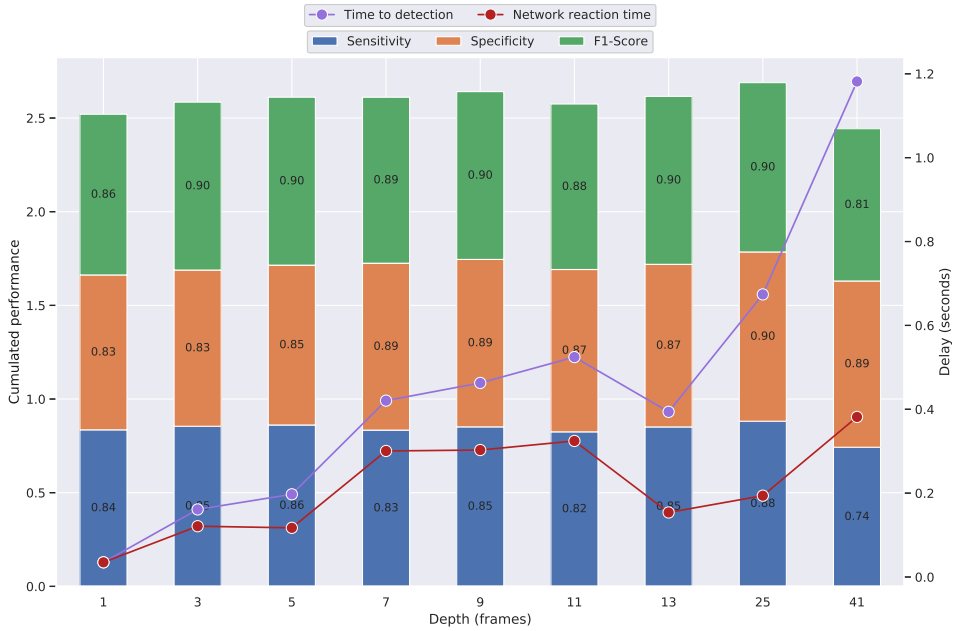


Fig. 7. Evaluation of the hybrid architecture on the *Video Dataset* when trained and tested with different input lengths, or temporal depths d. The stacked bars show the performance (left axis) and the line plot shows the average delay to first detect a polyp (right axis). (For interpretation of the references to colour, the reader is referred to the web version of this article.)

F1-Score was 90% for both the highest performing experiment ($d = 25$) and the one with the shortest inputs $d = 3$. Additionally, the performance could increase with larger temporal depths because of the nature of the data. Inherently, the *Video Dataset* is biased towards long polyp scenes. In this dataset, when a polyp is found, it stays in the scene for hundreds of frames for optical inspection. Similarly, there are long intervals in the procedures where there are no polyp appearances. In this sense, models trained for longer input samples would benefit from the stability of the data.

On the other hand, longer input temporal depths could negatively affect changes in the video, namely when a polyp appears or disappears from the scene. Arguably, the highest benefit of a polyp detection model resides in detecting polyps that would be missed otherwise, hence of short appearance in the colonoscopic video. It is subsequently important to introduce time to detection as a metric. As shown in Figure 7, the delay when first detecting a polyp tends to increase the longer the inputs are. Particularly the intrinsic delay from the model, the network reaction time, increases with longer input sizes up to a delay of 9.54 frames for $d = 41$. Additionally, longer inputs linearly increase the time to detection, leading to larger user-perceived delays (e.f. 29.54 frames for $d = 41$). It is worth mentioning that different sensitivity/specificity balances are obtained with different models, although they present a similar F1-Score. The time to detection is negatively correlated to the sensitivity, which might explain why experiments with 7, 9 and 11 frames have a higher delay.

All in all, benefits can be drawn from the hybrid architecture even with short input samples. Longer inputs introduce a delay when detecting polyps for the first time, but they also seem to bring a increase in performance. However, it would be necessary to test these models on prospective data to objectively evaluate its effect on missed polyps.

4. Discussion and Conclusion

In this paper a novel hybrid 2D/3D segmentation CNN architecture for polyp detection in colonoscopic videos was presented. It was shown that the hybrid network was able to encompass the benefits from a 2D architecture, namely successful spatial representation learning, transfer learning capabilities to pretrain from curated datasets of still images and generalisation abilities despite being trained on reduced amounts of data. The proposed architecture performed better than its 2D counterpart when trained with less data. It was additionally revealed that a great performance boost could be obtained from pretraining on still images when training the hybrid model on less data. This is particularly beneficial for clinical applications, where large video datasets are challenging to collect and hence still image data may be needed to provide strong and diverse representation of the spatial domain. Furthermore, the model was shown to generalise better to an external dataset and it exhibited similar performance to results published by the study that released said dataset. In the proposed method, the 3D segmentation seamlessly incorporated temporal correlation in the results encapsulating learning of spatio-temporal information from smaller video datasets. We obtained an increase across

all performance metrics, particularly in the temporal consistency of the results. Ablation studies comparing training the full model against solely updating the segmentation head demonstrated the benefit of the 3D decoding. Additional analysis assessed the effect of different temporal window sizes indicating that benefits could be obtained even with shorter inputs, with the added benefit of a lower delay to first detect a polyp.

Overall the hybrid architecture was validated on videos from 46 patients, including 25 unaltered, full-length negative procedures, offering an increase in performance along with higher quality segmentation potential. Moreover, the model was evaluated on a publicly available dataset containing 100 polyps, displaying promising results. All in all the hybrid network successfully harnessed temporal information from videos to handle short inconsistencies in the predictions hence showing a increased suitability for clinical translation. The model was evaluated in video data and user centred metrics were used such as time to detection, breaching the evaluation gap between "lab" and "hospital". The main limitation of the proposed solution is related to the introduced delay. The model has longer computation time than a 2D network, which is added to a longer intrinsic network reaction time and the fact that future frames are needed to compute a prediction on the current frame, yielding an overall time to first detection. Despite the presented architecture showing similar performance to the state-of-the-art on an external dataset, improvements in specificity are needed for a successful translation to the clinical setup. Although the issue can currently be minimised by post-processing techniques to reduce false positives, it leaves scope for improvement. Although the model was assessed on video data, bringing the evaluation closer to a real clinical setup, video datasets are biased towards longer sequences on found polyps and shorter appearances in the video from missed polyps. Future work includes evaluating the system on prospective data to fully analyse the effect of the delay of the hybrid architecture. Furthermore, it would be of interest to exploit the modularity of the Hybrid 2D/3D architecture and evaluate its performance with different backbones. Incorporation of depth and colon mapping information (Rau et al., 2019; Liu et al., 2018; Armin et al., 2018; Mathew et al., 2020; Itoh et al., 2021; Cheng et al., 2021) can be exploited to combine polyp detection with automated reporting tools, thus opening doors for polyp finding, referral and surveillance.

Acknowledgments. The work was supported by the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) [203145Z/16/Z]; Engineering and Physical Sciences Research Council (EPSRC) [EP/P027938/1, EP/R004080/1, EP/P012841/1]; The Royal Academy of Engineering [CiET1819 \2\36]; European Union's Horizon 2020 research and innovation programme under grant agreement No 863146; Work carried out under a programme of and funded by the European Space Agency, the view expressed herein can in no way be taken to reflect the official opinion of the European Space Agency.

References

Ahmad, O.F., Mori, Y., Misawa, M., Kudo, S.e., Anderson, J.T., Bernal, J., Berzin, T.M., Bisschops, R., Byrne, M.F., Chen, P.J., et al., 2020. Establishing key research questions for the implementation of artificial intelligence in colonoscopy-a modified delphi method. Endoscopy .

Ahmad, O.F., Soares, A.S., Mazomenos, E., Brandao, P., Vega, R., Seward, E., Stoyanov, D., Chand, M., Lovat, L.B., 2019a. Artificial intelligence and computer-aided diagnosis in colonoscopy: current evidence and future directions. The Lancet Gastroenterology & Hepatology 4, 71–80.

Ahmad, O.F., Stoyanov, D., Lovat, L.B., 2019b. Barriers and pitfalls for artificial intelligence in gastroenterology: ethical and regulatory issues. Techniques in Gastrointestinal Endoscopy , 150636.

Armin, M.A., Barnes, N., Khan, S., Liu, M., Grimpen, F., Salvado, O., 2018. Unsupervised learning of endoscopy video frames' correspondences from global and local transformation, in: OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis. Springer, pp. 108–117.

Bernal, J., Tajkbaksh, N., Sánchez, F.J., Matuszewski, B.J., Chen, H., Yu, L., Angermann, Q., Romain, O., Rustad, B., Balasingham, I., et al., 2017. Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge. IEEE Transactions on Medical Imaging 36, 1231–1249.

Brandao, P., Mazomenos, E., Ciuti, G., Caliò, R., Bianchi, F., Menciassi, A., Dario, P., Koulaouzidis, A., Arezzo, A., Stoyanov, D., 2017. Fully convolutional neural networks for polyp segmentation in colonoscopy, in: Medical Imaging 2017: Computer-Aided Diagnosis, International Society for Optics and Photonics. p. 101340F.

Brandao, P., Zisimopoulos, O., Mazomenos, E., Ciuti, G., Bernal, J., Visentini-Scarzanella, M., Menciassi, A., Dario, P., Koulaouzidis, A., Arezzo, A., et al., 2018. Towards a computed-aided diagnosis system in colonoscopy: automatic polyp segmentation using convolution neural networks. Journal of Medical Robotics Research 3, 1840002.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A., 2018. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a Cancer Journal for Clinicians 68, 394–424.

Carreira, J., Zisserman, A., 2017. Quo vadis, action recognition? a new model and the kinetics dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308.

Chahal, D., Byrne, M.F., 2020. A primer on artificial intelligence and its application to endoscopy. Gastrointestinal endoscopy 92, 813–820.

Cheng, K., Ma, Y., Sun, B., Li, Y., Chen, X., 2021. Depth estimation for colonoscopy images with self-supervised learning from videos, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 119–128.

Corley, D.A., Jensen, C.D., Marks, A.R., Zhao, W.K., Lee, J.K., Doubeni, C.A., Zauber, A.G., de Boer, J., Fireman, B.H., Schottinger, J.E., et al., 2014. Adenoma detection rate and risk of colorectal cancer and death. New England Journal of Medicine 370, 1298–1306.

Eelbode, T., Demedts, I., Bisschops, R., Roelandt, P., Hassan, C., Coron, E., Bhandari, P., Neumann, H., Pech, O., Repici, A., et al., 2019. Tu1931 incorporation of temporal information in a deep neural network improves performance level for automated polyp detection and delineation. Gastrointestinal Endoscopy 89, AB618–AB619.

Hassan, C., Wallace, M.B., Sharma, P., Maselli, R., Craviotto, V., Spadaccini, M., Repici, A., 2019. New artificial intelligence system: first validation study versus experienced endoscopists for colorectal polyp detection. Gut , gutjnl–2019.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Itoh, H., Oda, M., Mori, Y., Misawa, M., Kudo, S.E., Imai, K., Ito, S., Hotta, K., Takabatake, H., Mori, M., et al., 2021. Unsupervised colonoscopic depth estimation by domain translations with a lambertian-reflection keeping auxiliary task. International Journal of Computer Assisted Radiology and Surgery 16, 989–1001.

Itoh, H., Roth, H., Oda, M., Misawa, M., Mori, Y., Kudo, S.E., Mori, K., 2019. Stable polyp-scene classification via subsampling and residual learning from an imbalanced large dataset. Healthcare Technology Letters 6, 237–242.

Itoh, H., Roth, H.R., Lu, L., Oda, M., Misawa, M., Mori, Y., Kudo, S.e., Mori, K., 2018. Towards automated colonoscopy diagnosis: binary polyp size estimation via unsupervised depth learning, in: International Conference on Medical Image Computing and Computer-assisted Intervention, Springer. pp. 611–619.

Jha, D., H. Smedsrud, P., Riegler, M., Halvorsen, P., Johansen, D., de Lange,

T., D. Johansen, H., 2020. Kvasir-seg: A segmented polyp dataset, in: Proceedings of the International Conference on Multimedia Modeling (MMM), Springer. URL: https://datasets.simula.no/kvasir-seg/.

Kudo, S.e., Misawa, M., Mori, Y., Hotta, K., Ohtsuka, K., Ikematsu, H., Saito, Y., Takeda, K., Nakamura, H., Ichimasa, K., et al., 2020. Artificial intelligence-assisted system improves endoscopic identification of colorectal neoplasms. Clinical Gastroenterology and Hepatology 18, 1874–1881.

Leufkens, A., Van Oijen, M., Vleggaar, F., Siersema, P., 2012. Factors influencing the miss rate of polyps in a back-to-back colonoscopy study. Endoscopy 44, 470–475.

Liu, W.N., Zhang, Y.Y., Bian, X.Q., Wang, L.J., Yang, Q., Zhang, X.D., Huang, J., 2020. Study on detection rate of polyps and adenomas in artificial-intelligence-aided colonoscopy. Saudi journal of gastroenterology: official journal of the Saudi Gastroenterology Association 26, 13.

Liu, X., Sinha, A., Unberath, M., Ishii, M., Hager, G.D., Taylor, R.H., Reiter, A., 2018. Self-supervised learning for dense depth estimation in monocular endoscopy, in: OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis. Springer, pp. 128–138.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.

Ma, Y., Chen, X., Sun, B., 2020. Polyp detection in colonoscopy videos by bootstrapping via temporal consistency, in: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 1360–1363.

Mathew, S., Nadeem, S., Kumari, S., Kaufman, A., 2020. Augmenting colonoscopy using extended and directional cyclegan for lossy image translation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4696–4705.

Medical, P., 2019. Hoya group pentax medical cleared ce mark for discovery™, an ai assisted polyp detector. https://www.pentaxmedical.com/pentax/en/92/1/HOYA-Group-PENTAX-Medical-Cleared-CE-Mark-for-DISCOVERY-an-AI-Assisted-Polyp-Detector. Accessed: 23-030-2021.

Misawa, M., Kudo, S.e., Mori, Y., Cho, T., Kataoka, S., Yamauchi, A., Ogawa, Y., Maeda, Y., Takeda, K., Ichimasa, K., et al., 2018. Artificial intelligence-assisted polyp detection for colonoscopy: initial experience. Gastroenterology 154, 2027–2029.

Misawa, M., Kudo, S.e., Mori, Y., Hotta, K., Ohtsuka, K., Matsuda, T., Saito, S., Kudo, T., Baba, T., Ishida, F., et al., 2020. Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). Gastrointestinal Endoscopy .

Odin Vision, 2020. Caddie a new era of ai enabled endoscopy. https://odin-vision.com/wp-content/uploads/2020/11/OdinVision-CADDIE.pdf. Accessed: 13-5-2021.

Podlasek, J., Heesch, M., Podlasek, R., Kilisiński, W., Filip, R., 2021. Real-time deep learning-based colorectal polyp localization on clinical video footage achievable with a wide array of hardware configurations. Endoscopy International Open 9, E741–E748.

Poon, C.C., Jiang, Y., Zhang, R., Lo, W.W., Cheung, M.S., Yu, R., Zheng, Y., Wong, J.C., Liu, Q., Wong, S.H., et al., 2020. Ai-doscopist: a real-time deep-learning-based algorithm for localising polyps in colonoscopy videos with edge computing devices. NPJ Digital Medicine 3, 1–8.

Puyal, J.G.B., Bhatia, K.K., Brandao, P., Ahmad, O.F., Toth, D., Kader, R., Lovat, L., Mountney, P., Stoyanov, D., 2020. Endoscopic polyp segmentation using a hybrid 2d/3d cnn, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 295–305.

Qadir, H.A., Balasingham, I., Solhusvik, J., Bergsland, J., Aabakken, L., Shin, Y., 2019. Improving automatic polyp detection using cnn by exploiting temporal dependency in colonoscopy video. IEEE Journal of Biomedical and Health Informatics .

Rau, A., Edwards, P.E., Ahmad, O.F., Riordan, P., Janatka, M., Lovat, L.B., Stoyanov, D., 2019. Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. International Journal of Computer Assisted Radiology and Surgery 14, 1167–1176.

Repici, A., Badalamenti, M., Maselli, R., Correale, L., Radaelli, F., Rondonotti, E., Ferrara, E., Spadaccini, M., Alkandari, A., Fugazza, A., et al., 2020. Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. Gastroenterology 159, 512–520.

Rex, D.K., Johnson, D.A., Anderson, J.C., Schoenfeld, P.S., Burke, C.A., Inadomi, J.M., 2009. American college of gastroenterology guidelines for colorectal cancer screening 2008. American Journal of Gastroenterology 104, 739–750.

Su, J.R., Li, Z., Shao, X.J., Ji, C.R., Ji, R., Zhou, R.C., Li, G.C., Liu, G.Q., He, Y.S., Zuo, X.L., et al., 2020. Impact of a real-time automatic quality control system on colorectal polyp and adenoma detection: a prospective randomized controlled study (with videos). Gastrointestinal Endoscopy 91, 415–424.

Tajbakhsh, N., Gurudu, S.R., Liang, J., 2015. Automated polyp detection in colonoscopy videos using shape and context information. IEEE Transactions on Medical Imaging 35, 630–644.

Van Rijn, J.C., Reitsma, J.B., Stoker, J., Bossuyt, P.M., Van Deventer, S.J., Dekker, E., 2006. Polyp miss rate determined by tandem colonoscopy: a systematic review. American Journal of Gastroenterology 101, 343–350.

Wang, P., Li, L., Liu, P., Xiao, X., Song, Y., Zhang, D., Li, Y., Xu, G., Tu, M., Xiao, X., et al., 2018a. Mo1712 automatic polyp detection during colonoscopy increases adenoma detection: An interim analysis of a prospective randomized control study. Gastrointestinal Endoscopy 87, AB490–AB491.

Wang, P., Xiao, X., Brown, J.R.G., Berzin, T.M., Tu, M., Xiong, F., Hu, X., Liu, P., Song, Y., Zhang, D., et al., 2018b. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. Nature biomedical engineering 2, 741–748.

Weigt, J., Neumann, H., Repici, A., Hassan, C., 2020. Mit hilfe eines validierten polypendetektions-und charakterisierungssystems können unerfahrene untersucher expertennieveau erreichen. Zeitschrift für Gastroenterologie 58, P–097.

Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.A., 2016. Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos. IEEE Journal of Biomedical and Health Informatics 21, 65–75.

Zhang, P., Sun, X., Wang, D., Wang, X., Cao, Y., Liu, B., 2019. An efficient spatial-temporal polyp detection framework for colonoscopy video, in: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), IEEE. pp. 1252–1259.

Zhang, R., Zheng, Y., Poon, C.C., Shen, D., Lau, J.Y., 2018. Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker. Pattern recognition 83, 209–219.

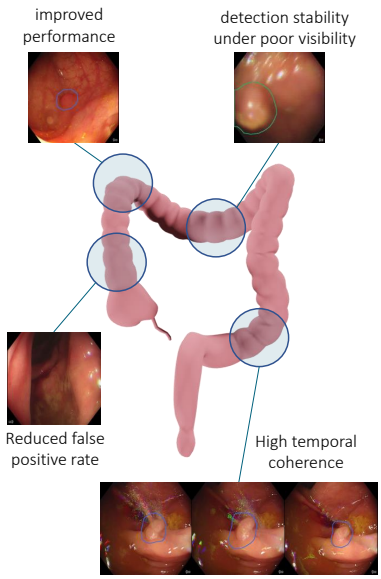
- Novel Hybrid 2D/3D convolutional neural network for video analysis
- Improves performance across the board when compared to 2D models
- High temporal coherence of the predictions
- Benefits of 3D model with generalisation capabilities and no need of large datasets
- Highly competitive results on external data

Polyp segmentation on video colonoscopy using a hybrid 2D/3D CNN

wellcome EPSRC centre interventional+ surgical sciences

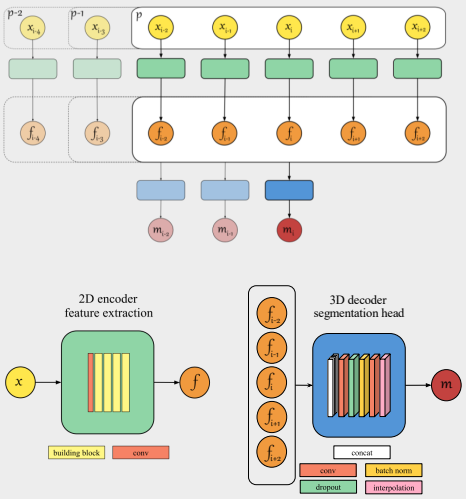
Odin Vision

Added benefits



Innovative architecture

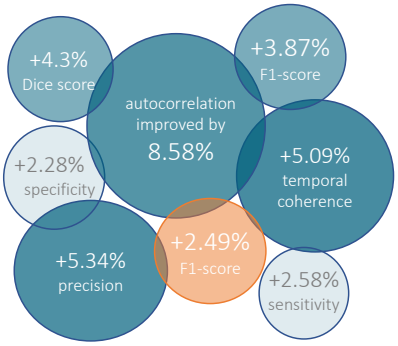
Hybrid 2D/3D CNN



Evaluated on

With respect to an FCN architecture

Video dataset	SUN database
563,526 frames	158,690 frames
53 polyps	100 polyps
46 patients	99 patients



Sensitivity and specificity comparable to the video dataset

Declaration of interests

☐ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☒The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

D.S, L.B.L and are involved with Odin Vision Ltd.
D.S is involved with Digital Surgery Ltd