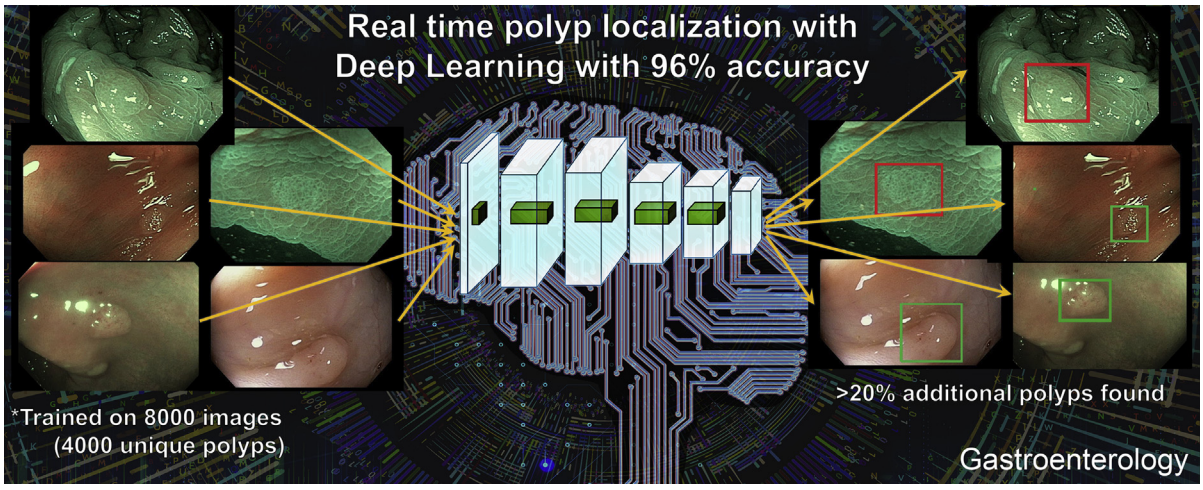
Gastroenterology 2018;155:1069–1078

Deep Learning Localizes and Identifies Polyps in Real Time With 96% Accuracy in Screening Colonoscopy

Gregor Urban,[1,2] Priyam Tripathi,[4] Talal Alkayali,[4,5] Mohit Mittal,[4] Farid Jalali,[4,5] William Karnes,[4,5] and Pierre Baldi[1,2,3]

[1]Department of Computer Science, University of California, Irvine, California; [2]Institute for Genomics and Bioinformatics, University of California, Irvine, California; [3]Center for Machine Learning and Intelligent Systems, University of California, Irvine, California; [4]Department of Medicine, University of California, Irvine, California; and [5]H.H. Chao Comprehensive Digestive Disease Center, University of California, Irvine, California



BACKGROUND & AIMS: The benefit of colonoscopy for colorectal cancer prevention depends on the adenoma detection rate (ADR). The ADR should reflect the adenoma prevalence rate, which is estimated to be higher than 50% in the screening-age population. However, the ADR by colonoscopists varies from 7% to 53%. It is estimated that every 1% increase in ADR lowers the risk of interval colorectal cancers by 3%–6%. New strategies are needed to increase the ADR during colonoscopy. We tested the ability of computer-assisted image analysis using convolutional neural networks (CNNs; a deep learning model for image analysis) to improve polyp detection, a surrogate of ADR. **METHODS:** We designed and trained deep CNNs to detect polyps using a diverse and representative set of 8,641 hand-labeled images from screening colonoscopies collected from more than 2000 patients. We tested the models on 20 colonoscopy videos with a total duration of 5 hours. Expert colonoscopists were asked to identify all polyps in 9 de-identified colonoscopy videos, which were selected from archived video studies, with or without benefit of the CNN overlay. Their findings were compared with those of the CNN using CNN-assisted expert review as the reference. **RESULTS:** When tested on manually labeled images, the CNN identified polyps with an area under the receiver operating characteristic curve of 0.991 and an accuracy of 96.4%. In the analysis of colonoscopy videos in which 28 polyps were removed, 4 expert reviewers identified 8 additional polyps without CNN assistance that had not been removed and identified an additional 17 polyps with CNN assistance (45 in total). All polyps removed and identified by expert review were detected by the CNN. The CNN had a false-positive rate of 7%. **CONCLUSION:** In a set of 8,641 colonoscopy images containing 4,088 unique polyps, the CNN identified polyps with a cross-validation accuracy of 96.4% and an area under the receiver operating characteristic curve of 0.991. The CNN system detected and localized polyps well within real-time constraints using an ordinary desktop machine with a contemporary graphics processing unit. This system could increase the ADR and decrease interval colorectal cancers but requires validation in large multicenter trials.

Keywords: Machine Learning; Convolutional Neural Networks; Colorectal Cancer Prevention; Adenoma Detection Rate Improving Technology.

Abbreviations used in this paper: **ADR, adenoma detection rate; AUC, area under the curve; CNN, convolutional neural network; CRC, colorectal cancer; FPR, false-positive rate; NBI, narrow-band imaging; NPI, not pre-initialized; PI, pre-initialized; WLE, white light endoscopy.**

Most current article

0016-5085/$36.00
https://doi.org/10.1053/j.gastro.2018.06.037

WHAT YOU NEED TO KNOW

BACKGROUND AND CONTEXT

The benefit of colonoscopy for colorectal cancer prevention depends on the adenoma detection rate (ADR). New strategies are needed to increase the ADR during colonoscopy.

NEW FINDINGS

A system of convolutional neural networks (CNN) called Deep Learning was able to process colonoscopy images at high speed in real time, identifying polyps with a cross-validation accuracy of 96.4% and ROC-AUC value of 0.991.

LIMITATIONS

Possible effects of the CNN on inspection behavior by colonoscopists are not known. The anonymized videos excluded information about patient history. CNN performance may vary by indication (screening vs surveillance).

IMPACT

This technology may assist colonoscopists in finding precancerous polyps in real-time and with high accuracy.

Colorectal cancer (CRC) is the second leading cause of cancer-related death in the United States.[1] CRC arises from precancerous polyps[2] with a mean dwell time of at least 10 years.[3] The National Polyp Study showed that 70%–90% of CRCs are preventable with regular colonoscopies and removal of polyps.[4] Seven percent of 9% of CRCs occur despite being up to date with colonoscopy.[5] It is estimated that 85% of these "interval cancers" are due to missed polyps or incompletely removed polyps during colonoscopy.[6]

The prevalence of precancerous polyps in the screening population older than 50 years is estimated to be at least 50%.[7] Adenomas are the most prevalent precancerous polyp. The adenoma detection rate (ADR; percentage of screening colonoscopies with $\geq$1 adenoma found) is a measure of a colonoscopist's ability to find adenomas. Ideally, the ADR should reflect adenoma prevalence. Unfortunately, the ADR varies widely (7%–53%) among colonoscopists performing screening colonoscopies.[8] In tandem colonoscopies, 22%–28% of polyps and 20%–24% of adenomas were missed[7] and CRC had a diagnostic miss rate of 5%.[9] The ADR is dependent on a colonoscopist's level of training, time spent, and technique used during withdrawal, preparation quality, and other colonoscopist- and procedure-dependent factors.[10] A large Kaiser Permanente study showed that for each 1% increase in the ADR, the interval cancer rate was decreased by 3%.[8] A subsequent study with nearly 1 million person-years of follow-ups in Poland showed a 6% decrease in interval cancer rates for each 1% increase in the ADR.[11] This study also showed an 82% decrease in interval cancer rates among colonoscopists that improved their ADRs to the top quintile. Not surprisingly, the ADR currently is a key quality measure reportable in the United States to the Centers for Medicare and Medicaid and is tied to reimbursement under the Medicare Access and CHIP Reauthorization Act of 2015 and the Merit Based Incentive Payments System beginning in the 2017.[12]

Several novel technologies have been developed to improve the ADR, including enhanced optics (resolution, zoom and wide angle, chromoendoscopy, digital autofluorescence, extra lenses for side and forward views) and attachments and modifications to aid view behind and between folds, including cap-assisted techniques and a balloon-assisted device.[13] Extra-wide angle colonoscopes and multi-camera systems initially showed promise to increase the ADR compared with standard forward-facing camera systems.[13] However, a recent meta-analysis and large randomized study showed no difference in the ADR compared with standard forward-viewing colonoscopy.[14] A review of 5 studies on the effect of high-definition colonoscopes on the ADR showed conflicting evidence,[13] with 1 study concluding that the ADR is improved only for endoscopists with a low ADR (<20%).[15] Similarly, most studies on digital chromoendoscopy, specifically narrow-band imaging (NBI), have found that NBI does not improve the ADR compared with white light imaging.[13] Evidence suggests positive effects of autofluorescence, but it is associated with added expense and poor image resolution.[13]

Computer-assisted image analysis has the potential to further aid adenoma detection but has remained underdeveloped. A notable benefit of such a system is that no alteration of the colonoscope or procedure is necessary.

Deep learning has been successfully applied to many areas of science and technology,[16] such as computer vision,[17–21] speech recognition,[22] natural language processing,[23] games,[24,25] particle physics,[26,27] organic chemistry,[28] and biology,[29–34] to name just a few areas and examples. A convolutional neural network (CNN) is a type of deep learning model that is highly effective at performing image analysis.

Ideally, a polyp-detection assistance module should have a sensitivity of 1 (or close to it) to avoid false-negative results, but this comes at the cost of an increased false-positive rate (FPR) when the area under the curve (AUC; Performance Evaluation and Metrics section in the Supplement) is not close to 1. A large FPR, even with perfect sensitivity, diminishes the benefits of an assistance system when user desensitization comes into play. A polyp-detection module also must process images at a minimum of 30 frames per second to be applicable during colonoscopy. Therefore, surmounting the constraints of accuracy and processing speed were our primary goals.

Methods

Convolutional Neural Networks

We trained different CNN architectures in this study, including models with weights initialized by training on the ImageNet data corpus,[35] before refining the weights in our dataset. All trained CNNs consisted of the same fundamental building blocks, including (1) convolutional layers, (2) fully

connected layers, (3) maximum or average pooling, (4) nonlinear activation functions, and (5), optionally, batch normalization operations[36] and skip connections.[19,37]

We followed each convolutional layer by the rectified linear (ReLu) activation function. The last hidden layer of the models was densely connected to the output units. For regression problems (localization), we optimized the L2 loss with linear output units. For classification (detection), we used softmax output units and optimized the Kullback–Leibler divergence. An overview of the different neural network layer types is presented in the Neural Network Architectures section in the Supplementary Material while Supplementary Table 1 presents the used CNN architectures.

All experiments were implemented using the Keras[38] and Tensorflow[39] software libraries.

Model Regularization

We used established techniques to decrease overfitting when training neural networks. We applied dropout[40,41] with a rate of 0.5 to the input of the first and second fully connected layers in all models. Prior studies have found that data augmentation improves deep learning performance,[42] a process of synthetically generating additional training examples by using random image transformations, including rotations and mirroring of the input images during the training process. Doing so forces the model to learn to become invariant to these transformations. We used random horizontal and vertical mirroring, rotations in the full range of 0°–90°, and shearing. Another technique we used to decrease overfitting was "early stopping," in which a small subset of the training set is reserved exclusively for monitoring the CNN's accuracy during training and the weights of the network at the point of best performance are saved, as opposed to the weights obtained at the end of training.

Experiments

Datasets and Preprocessing

Five different datasets were used for training and/or evaluating the deep learning models presented in this work: (1) the general-purpose computer-vision ImageNet challenge[35] dataset was implicitly used to pre-train the model weights; (2) 8,641 hand-selected colonoscopy images from more than 2,000 patients were used to avoid a possible intrapatient polyp similarity bias; (3) a separately collected dataset of 1,330 colonoscopy images from different patients; (4) 9 colonoscopy videos; (5) a combined dataset consisting of the 8,641 images and 44,947 image frames extracted from the 9 videos; and (6) a separate dataset of 11 deliberately more "challenging" colonoscopy videos. All colonoscopy images and videos were recorded using state-of-the-art Olympus PCF-H190 colonoscopes (Olympus Corp, Center Valley, PA) with a resolution of 1,280 × 1,024 pixels and NBI capability. The combined duration of all 20 videos was approximately 5 hours (~500,000 frames), with significant variation in the number of polyps per video (See Supplementary Tables 2–4 and especially Supplementary Table 5 in this context).

1. The ImageNet challenge dataset contains 1.2 million natural images of objects, such as boats, cars, and dogs, but no medical images. We reasoned that many fundamental features learnable on this dataset could be transferable to the task of detecting polyps and thus used it to pre-initialize the weights of some of our deep neural networks to test this hypothesis.
2. The set of 8641 colonoscopy images contained 4,088 images of unique polyps of all sizes and morphologies and 4553 images without polyps (ie, the dataset was almost perfectly balanced; Figure 1). The dataset included white light and NBI images (Figure 1) and covered all portions of the colorectum, including retro-views in the rectum and cecum, appendiceal orifice, and ileocecal valve. The total number of NBI images was 840, with the remaining 7801 obtained in white light endoscopy (WLE) conditions. We deliberately and randomly included features such as forceps, snares, cuff devices, debris, melanosis coli, and diverticula in polyp- and non–polyp-containing images in a balanced fashion to prevent the machine learning system from associating the appearance of tools with the presence of polyps. The images were stored at a resolution of 640 × 480 pixels. Locations and dimensions of bounding boxes were recorded for images containing polyps by a team of colonoscopists (fellows and faculty at the University of California–Irvine with an ADR > 45% and >100 procedures).
3. A separate set of 1330 colonoscopy images (672 unique polyp and 658 non-polyp images) was collected from different patients.
4. Colonoscopy videos were recorded and evaluated at their original resolution of 1280 × 1024 pixels. The first set of 9 videos was selected randomly from archived video studies.
5. The larger dataset was obtained by augmenting the original set of 8641 images with 44,947 image frames selected from the 9 videos that were labeled as part of the first validation study. Because consecutive frames are highly correlated, we selected every eighth image frame that contained no polyp and every fourth image frame containing a polyp (resulting in 13,292 polyp frames and 31,655 non-polyp frames).
6. Colonoscopy procedures of the second set of 11 videos were performed by a highly skilled colonoscopist (ADR ≥ 50%) and contained segments for which the scope was deliberately withdrawn without closing in on already identified polyps to mimic a missed-polyp scenario. This set of videos was used only for validation purposes in our experiments and never for training.

All images from the different datasets were preprocessed identically before being passed to the machine learning models. As a first step, the individual frames were rescaled to a fixed size of 224 × 224 (unless noted otherwise). Then, the values of the pixels in each frame were normalized to be unit normally distributed by subtracting the mean pixel value from all pixels in the frame and dividing the resulting values by the standard deviation

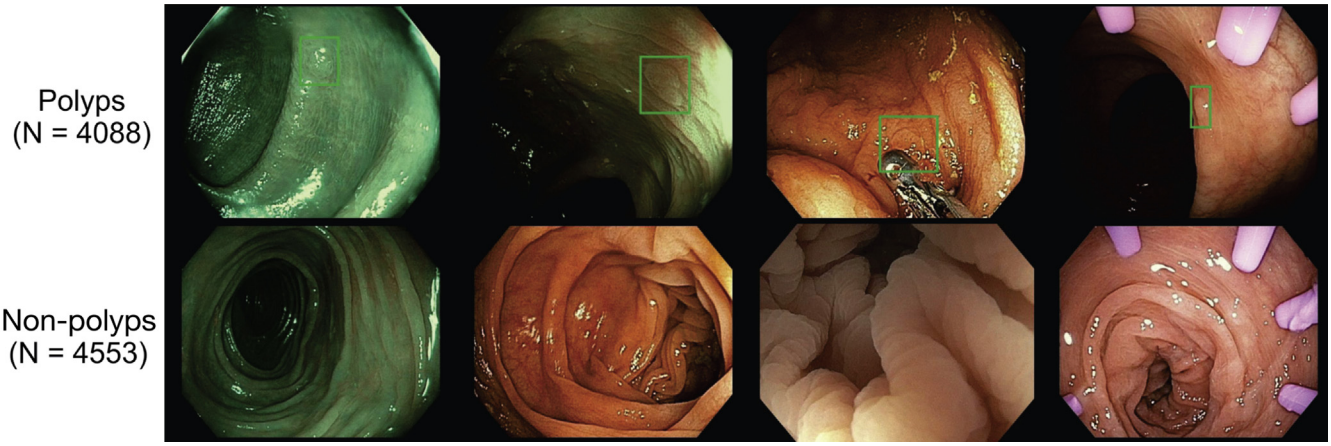


Figure 1. Examples of dataset. (*Top row*) Images containing a polyp with a superimposed bounding box. (*Bottom row*) Non-polyp images. Three pictures on the *left* were taken using NBI and 3 pictures on the *right* include tools (eg, biopsy forceps, cuff devices, etc) that are commonly used in screening colonoscopy procedures.

measured across all pixels. This preprocessing and normalization approach allowed us to apply the same neural network to data from different sources, with different resolutions, without requiring further adjustments.

Neural Network Architectures and Training

A detailed description of the deep neural network architectures, training methods, and algorithms is presented in the Architectures, Training, and Evaluation section in the Supplement. A short summary is given below.

We trained and evaluated polyp-detection and -localization models separately for clarity and to exclude confounding factors. They were architecturally identical except for the final layer, which performed binary classification (detection) or regression (localization). We implemented localization by predicting the size and location of a bounding box that tightly enclosed any identified polyps. This allowed us to build CNNs that could operate in real time as opposed to prior attempts, based on individual pixel classification, that struggled to operate in real time.[43] We tested 3 variations for training the polyp localization model: (1) optimizing the size and location with the mean-squared error loss (L2); (2) optimizing the Dice loss, which directly maximizes the overlap between the predicted bounding box and the ground truth; and (3) a variation of the "you only look once"[20] algorithm in which the CNN produces and aggregates typically more than 49 individual weighted predictions of polyp size and location in a single forward pass; we refer to it as the "internal ensemble." All detection and localization variants had almost identical run-time complexity (<1% difference).

We experimented with several different CNN architectures (Supplementary Table 1) that fell into 1 of 2 categories: those initialized with random weights (denoted NPI for "not pre-initialized") and those pre-initialized (PI) with weights obtained by training the corresponding model on the ImageNet Challenge data of natural images. Within the class of PI architectures, we used architectures VGG16,[18] VGG19,[18] and ResNet50.[19] All CNN architectures (NPI and PI) were trained using the colonoscopy data. All experiments were performed using modern TITAN X (Pascal) graphics processing units (NVIDIA, Santa Clara, CA) with 12 GB of RAM and a processing power of 11 TFLOPS.

Training and Testing

Throughout the experiments, we used multiple splits of the same dataset (cross-validation) or trained on 1 dataset and tested the model on a completely different dataset. For early stopping and hyper-parameter optimization, we always set aside a small subset of the training set for monitoring the model's performance. The main experiments were:

- Cross-validation on the 8641 images
- Training on the 8641 images and testing on the 9 videos, 11 videos, and independent dataset
- Training on the 8641 images and 9 videos and testing on the 11 videos and independent dataset

In most cases we use models pre-trained on ImageNet.

Colonoscopy Video Study With Expert Colonoscopists

Three expert colonoscopists (ADR $\geq$ 50%) were tasked to identify all polyps in 9 de-identified colonoscopy videos selected from the archived video studies without benefit of the CNN overlay. Experts recorded the first and last frames in which they believed they encountered a polyp in the videos. Their "polyp encounters" were combined by consensus. We filtered the CNN predictions for polyps by requiring at least 8 contiguous video frames with greater than 40% probability for polyp presence as predicted by the PI-CNN2 model alone. The decision to filter CNN predictions in blocks of 8 frames yielded a balanced sensitivity and specificity (see Supplementary Figure 1 for an analysis of different block sizes). This optimal setting might shift for a different training dataset or CNN model.

We generated CNN-overlaid videos by superimposing a small green box on each frame where a polyp was detected

with greater than 95% predicted probability at the location and with dimensions that were predicted by our polyp-localization CNN.

A senior expert (ADR $\geq$ 50%, >20,000 colonoscopies) was tasked to review the CNN-overlaid videos and assign the uniqueness of each polyp and the confidence level of true polyp presence (high vs low; Figure 2). Contingency analysis compared the number of agreements and disagreements on individual video frames between post-processed CNN predictions alone and CNN-assisted expert review. We repeated this study using a second set of 11 more challenging videos (Datasets and Preprocessing section).

Results

Polyp Detection

Polyp detection results are presented in Table 1 (see Supplementary Figure 2 for the corresponding ROC plots). The first 2 rows (NPI-CNN1 and -2) correspond to models that were trained starting from random weights and these obtained accuracies that were comparable to previously published state-of-the-art polyp classification models.[44]

Networks pre-initialized from prior training on the large ImageNet dataset of natural images surpassed those starting from random weights by a significant margin (PI-CNN1–3), despite meticulous hyper-parameter optimization of all models. We also observed that the scores of the PI ImageNet models were surprisingly similar for the different architectures.

At a sensitivity level (true-positive rate) of 90%, the best model (PI-CNN2) had an FPR of 0.5%; at a sensitivity of 95.2%, the FPR was 1.6%; and at 97.1% sensitivity, the FPR increased to 6.5%. Thus, there was a reasonably large range of high sensitivities at which the number of expected false-positive alerts remained very low.

Nonpolypoid (flat and depressed) polyps are challenging to detect compared with polypoid polyps and were often neglected until their potential to cause CRC worldwide (beyond Japan only) was discovered.[45] To investigate whether the CNN could correctly classify all polyps irrespective of their morphology, we reviewed a random subset of 1578 true-positive polyp predictions (of 3860) and all 228 false-negative polyp predictions from the validation set of 8641 images. We categorized them using the Paris classification and their estimated size. The results are presented in Table 2 and show that 381 nonpolypoid lesions (IIa, IIb, and IIc) and 678 polypoid polyps (Ip and Is) were present in this subset. The CNN missed 12% of polypoid polyps (84 of 678) and 11% of nonpolypoid lesions (41 of 381) in this biased subset. Based on this similarity, we conclude that the CNN can detect either polyp type equally well. Furthermore, if we correct for the sampling bias by including all the remaining true-positive values, then the percentage of missed polyps decreases to approximately 5%.

All previously discussed results were obtained with models operating on inputs scaled to 224 $\times$ 224 pixels, which is the native resolution for which VGG16, VGG19, and ResNet50 were designed. We also optimized and trained the models at a resolution of 480 $\times$ 480 pixels to test whether they would be limited by the lower resolution. In a 7-fold cross-validation experiment, the networks pre-initialized with weights from the VGG16, VGG19, and ResNet50 models yielded test accuracies of 96.4%, 96.1%, and 96.4% and AUCs of 0.990, 0.991, and 0.992, respectively. These results were almost identical (up to noise) to those obtained with a lower resolution of 224 $\times$ 224 pixels but more than double the computational cost (processing time).

The VGG-19–based CNN trained on the 8,641 images was tested on the independent dataset of 1,330 images, where it achieved a test accuracy of 96.4% and an AUC of 0.974. This accuracy was identical to the accuracy obtained by cross-validation analysis on the 8641 colonoscopy images, further confirming that intrapatient polyp similarity does not present a notable bias.

Polyp Localization

We trained models on the polyp-localization task on the data subset containing only a single polyp per frame, which represents the vast majority of samples with polyps.

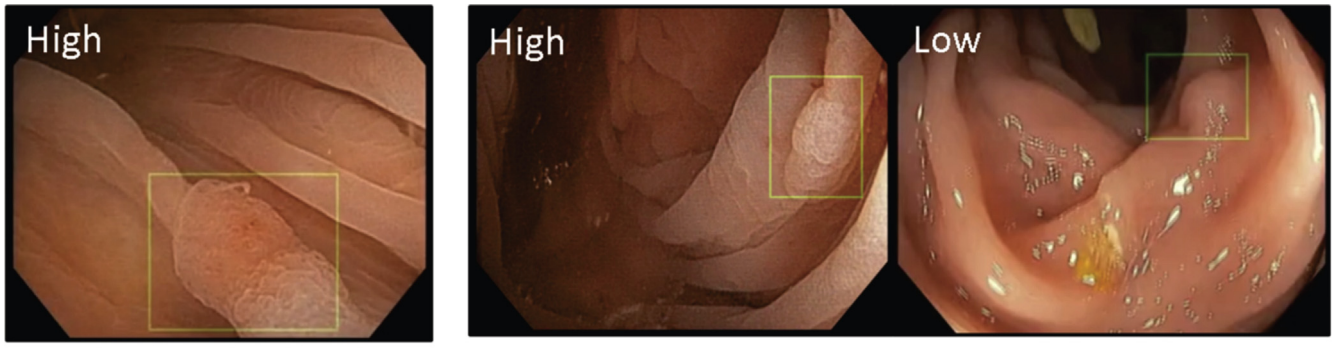


Figure 2. Representative frame shots of CNN-overlaid colonoscopy videos. Presence of a *green box* indicates that a polyp is detected with greater than 95% confidence by our CNN polyp localization model; the location and size of the box are predictions of the CNN model. Expert confidence that the box contained a true polyp is shown in the *upper left* of the images (video collages of CNN localization predictions available at: http://www.igb.uci.edu/colonoscopy/AI_for_GI.html).

Table 1. Summary of Polyp Classification Results for Architectures (Supplementary Table 1) Obtained by 7-Fold Cross-Validation on 8,641 Colonoscopy Images

Model	Initial weights	Accuracy	AUC	Sensitivity at 5% FPR	Sensitivity at 1% FPR
NPI-CNN1	—	91.9 ± 0.2%	0.970 ± 0.002	88.1%	65.4%
NPI-CNN2	—	91.0 ± 0.4%	0.966 ± 0.002	86.2%	60.6%
PI-CNN1	VGG16	95.9 ± 0.3%	0.990 ± 0.001	96.9%	87.8%
PI-CNN2	VGG19	96.4 ± 0.3%	0.991 ± 0.001	96.9%	88.1%
PI-CNN3	ResNet50	96.1 ± 0.1%	0.990 ± 0.001	96.8%	88.0%

NOTE. The sensitivity (true-positive rate) is given at false-positive rates of 5% and 1% (ie, a specificity of 95% and 99%, respectively), because sensitivity and specificity are interdependent values.
FPR, false-positive rate.

The test set results, as presented in Table 3, showed that the PI ImageNet CNNs (PI-CNN1 and -2) performed significantly better at localizing polyps than the randomly initialized neural network, which is consistent with our findings for polyp-presence detection (Polyp Detection section).

Neither of the 2 loss functions (L2 vs Dice) showed a consistent advantage over the other.

Further, we found that the "internal ensemble" was noticeably better than the other approaches by improving the Dice coefficient from 0.79 to 0.83 for the best model (PI-CNN2 and VGG19).

To provide a sense of scale, in previously published literature,[43] a Dice score of 0.55 was obtained on a polyp-segmentation task using a different dataset.

Colonoscopy Video Study With Expert Colonoscopists

Of 9 colonoscopy videos that we considered in this study, 36 polyps were identified by 3 experts reviewing unaltered videos and 45 were identified by reviewing CNN-overlaid videos. Only 28 of 45 polyps were removed by the original colonoscopists (Table 4). No unique polyps were missed by the CNN. Of the 9 additional polyps found with CNN assistance, confidence was high for 3 and low for 6 (cf Figure 2). The sensitivity and specificity of CNN predictions compared with expert review of CNN-overlaid videos (based on single frames) were 0.93 and 0.93, respectively ($P < .00001$ by χ^2 test; Supplementary Table 3). Rare false-negative results were enriched with views of distant and field-edge polyps. False-positive results were enriched with near-field collapsed mucosa, debris, suction marks, NBI, and polypectomy sites. Tables 4 and 5 present polyp-level summaries of the results of the video study.

The second set of 11 videos contained 73 unique polyps found by expert review with and without CNN assistance. The CNN trained on the 8,641 images identified 68 of 73 polyps at a frame-by-frame FPR of 7% compared with expert labels of the videos. Fine tuning the CNN on the labeled frames from the first video study enabled the CNN to identify 67 of 73 polyps at an FPR of 5% (or, on a more sensitive setting, 72 of 73 polyps at an FPR of 12%) in this second set of videos. The additional training samples from the videos of the first study noticeably helped decrease the number of false-positive detections, most likely because of the great abundance and variety of random artifacts such as water, air bubbles, fecal matter, and low-quality and blurry frames from quick movement. This suggests that using additional training data could lead to further improvements.

The Supplement contains a detailed breakdown of the polyps found in the 20 videos by the experts, including their location, size, morphology, and other details.

Table 2. Categorization of a Random Subset of 1,578 True-Positive and All 228 False-Negative Polyp CNN Predictions on the Test Set of 4,088 Unique Polyps, Categorized by Size and Paris Classification

	True positives		False negatives	
	≤1 cm	>1 cm	≤1 cm	>1 cm
Diameter (≤3 mm)	644	—	103	—
Ip	37	25	8	6
Is	487	45	68	2
IIa	246	37	36	4
IIb	34	15	1	0
IIc	4	4	0	0

NOTE. Results were obtained by 7-fold cross-validation on the 8,641 colonoscopy images. All polyps larger than 3 mm are categorized by the Paris classification scheme. The CNN performs equally well at detecting nonpolypoid lesions (IIa–c) and polypoid polyps (Ip and Is).

Table 3. Summary of Polyp Localization Results for a Subset of Architectures (Supplementary Table 1) Obtained by 7-Fold Cross-Validation on 8,641 Colonoscopy Images

Model	L2 regression (Dice)	Dice score optimization (Dice)	"Internal ensemble" regression (Dice)
NPI-CNN1	0.63 ± 0.01	0.681 ± 0.002	0.731 ± 0.006
PI-CNN1	0.77 ± 0.01	0.76 ± 0.01	0.823 ± 0.003
PI-CNN2	0.79 ± 0.01	0.784 ± 0.004	0.827 ± 0.003

NOTE. Standard deviation of mean estimate from the cross-validation is shown.

Table 4. Unique Polyps Found and Removed During Colonoscopy, Found by Expert Review, and Found by CNN-Assisted Expert Review of 9 Videos

Polyp size (*mm*)	Original colonoscopist (polyps removed)	Expert review	CNN-assisted review
1–3	12	19	24
4–6	12	13	16
7–9	0	0	1
>10	4	4	4
Total polyps found	28	36	45

NOTE. The VGG-19–based CNN was trained on the 8,641 colonoscopy images and applied to the 9 videos without further adaptation.

Additional Experiments

We performed further experiments to test (1) whether a system trained on WLE and NBI could perform well on colonoscopies without NBI capability and (2) whether a system trained on WLE data could perform even better on WLE-only colonoscopies than a system trained using NBI and WLE. We retrained the same VGG-19–based CNN on NBI-only, WLE-only, and subsets of the 8,641 colonoscopy images in 7-fold cross-validation analyses. Training and testing the CNN on WLE-only data resulted in a cross-validation test accuracy of 96.1% and an AUC of 0.991, whereas training and testing on NBI-only data yielded an accuracy of 92.9% and an AUC of 0.970. In these cases, this was worse than what the same CNN achieved when trained on WLE plus NBI data: an accuracy of 96.4% with an AUC of 0.992 on WLE-only data and an accuracy of 94.8% and an AUC of 0.988 on NBI-only data. The test accuracy on NBI images was consistently worse than on WLE data, but this can be explained by the significantly smaller amount of NBI training data (840 NBI images vs 7,801 WLE images). In summary, we found that it was beneficial to train the CNN on NBI plus WLE images, because this increased the total amount of training data, resulting in a synergistic knowledge "transfer" between the 2 modalities.

Table 5. Analysis of VGG-19–Based CNN Tested on All 20 Videos After Training on 8,641 Colonoscopy Images

	9 Videos	11 "Challenging" videos[a]
Total polyps found	45	68
Total polyps missed	0	5
Total false positives	81	46
Average delay to detection (s)[b]	0.2 ± 0.1	1.3 ± 0.3

NOTE. CNN-assisted expert review annotations were used as reference.
[a]"Challenging videos" were produced by an expert colonoscopists who performed "flybys" that included but avoided inspection of known polyps to mimic missed-polyp scenarios.
[b]Delay to detection of a polyp is the time span from a polyp entering the field of view of the camera for the first time to the CNN producing its first positive polyp prediction. All false-positive findings with duration of at least 1 second are counted. A frame-by-frame analysis of the video study is presented in Supplementary Table 3.

Discussion

An early application of computer-assisted polyp detection used traditional non-learning–based computer-vision engineering methods and achieved an AUC of 0.98 for detecting a limited class of polyp types[43] but could not work in real time, requiring 19 seconds to process a single frame. Of 8 submissions to the MICCAI 2015 Endoscopic Vision Challenge for polyp detection, none could operate in real time, and the most accurate (CUMED) had a detection accuracy of 89% tested across 18,092 video frames.[43] Other CNNs applied to the task of polyp detection have been limited by small databases of polyp images and videos. An implementation operating on multiple small sub-patches of images reached a classification accuracy of 91%,[44] whereas another approach using 3 convolutional layers operating on heavily subsampled images of 32 × 32 pixels obtained an AUC of 0.86, a sensitivity of 86%, and a specificity of 85%.[46]

In this study we trained state-of-the-art CNNs, pre-initialized on millions of labeled natural images (ImageNet), on a dataset of more than 8,000 labeled colonoscopy images from more than 2,000 patients. For the first time, these models could identify and locate polyps in real time and simultaneously achieve high accuracy and AUC. The highest performing model could process 98 images per second (10 ms per frame) for polyp detection and localization when using a modern consumer-grade graphics processing unit. This is approximately 4 times faster than required to implement real-time assistance software, because commonly used video encodings are standardized to 25 or 30 frames per second (PAL and NTSC standards). In comparison, the fastest polyp localization model presented in prior work,[47] among 8, could process only 7 frames per second and the slowest one could process merely 0.1 frame per second. A main contribution to the speed of our system comes from the choice of locating polyps through bounding boxes rather than unnecessarily precise polyp boundaries.

For resolution, although all human assessments of videos were performed at high resolution, we found that a 224- × 224-pixel resolution was sufficient for the CNNs and virtually indistinguishable from a 480- × 480-pixel resolution. We believe that future advances in computer hardware and machine learning will allow the use of even larger input fields while preserving real-time capabilities and potentially improving detection accuracy.

Our feasibility study of 9 colonoscopy videos, reviewed by expert colonoscopists with and without the aid of a CNN overlay, showed that the model identified all polyps and aided discovery of additional polyps with relatively a low burden of false-positive results. In a second study, with 11 purposefully difficult colonoscopy videos, recorded by a senior colonoscopist, featuring "flyby" scenarios without closing in on previously found polyps during withdrawal, the CNN alone identified 67 of 73 unique polyps, with an

average of fewer than 5 false-positive results per video. Missed polyps were located in "flyby" segments of the video, suggesting that CNN assistance cannot compensate for hurried withdrawal and poor inspection technique. Polyp morphology did not play a role in CNN performance, that is, nonpolypoid lesions were not missed by the CNN more often than polypoid polyps.

Our feasibility study suggests that CNN assistance during live colonoscopy will result in fewer missed polyps. However, extrapolation to real-time use is limited by several factors, including unknown effects of the CNN on inspection behavior by the colonoscopist. Another limitation derives from the anonymized and de-identified nature of the videos, which excluded information about the indications for colonoscopy or the histology of polyps. CNN performance can vary by indication (screening vs surveillance).

Polyp histology is especially relevant for added time and pathology costs. Time spent for polypectomy has "added value" whenever a true-positive result is precancerous, malignant, or relevant for calculating surveillance interval cancer and/or ADR. However, if use of the CNN results in polypectomies of clinically irrelevant lesions, the added time and unnecessary pathology costs would be unacceptable. Future randomized studies could directly address the overall value (quality vs cost) of the CNN by examining its effects on colonoscopy time, pathology costs, ADR, polyps per procedure, surveillance-relevant polyps per procedure, and surveillance-irrelevant polyps per procedure (normal, lymphoid aggregates, etc).

Live use of the CNN can lengthen colonoscopy procedure times owing to second looks at false-positive findings and additional polypectomies. Time to assess a false-positive finding will likely average less than 5 seconds at an estimated FPR of fewer than 8 per colonoscopy. This relatively minor time cost could be lowered with further optimization of detection accuracy (eg, more training data), user interface (eg, color selection, sound effects), and simultaneous running of accurate optical pathology artifical intelligence algorithms.

Although our results were obtained using Olympus endoscopes, which have a 70% endoscope market share,[48] we expect that the proposed method will work with endoscopes from other vendors with little additional tuning of the algorithm. This is consistent with hundreds of experiments reported in the literature on "transfer learning" and our extensive experience with other biomedical imaging problems.[30–32]

Our proposed method shows great promise in helping to close the gap between ADR and true adenoma prevalence, especially for the colonoscopist with a low ADR. By meeting the constraints of high accuracy and real-time performance using standard personal computers, this is the first reported polyp-detection artificial intelligence application ready for real-time validation studies.

Conclusion

We built a state-of-the-art polyp detection and localization model using deep learning that is easily capable of operating in real-time conditions (processing 1 frame in 10 ms). We detected the presence of polyps in a frame with an accuracy of 96.4% and an AUC of 0.991 using a CNN that was first trained on the ImageNet corpus of natural images and then retrained on our polyp database. A small adaptation of the model enabled it to localize polyps to within a bounding box with a state-of-the-art Dice/F1 score of 0.83 with a processing time of only 10 ms per frame. When overlaid on colonoscopy videos, the algorithm identified all polyps found by expert viewers (ADR $>$ 50%) and additional polyps missed at expert review of non-overlaid videos. Therefore, we believe that, when running live during colonoscopy, this model will prompt more careful inspection and discovery of additional polyps. Thus, it is well positioned for validation in prospective trials to test effectiveness for improving ADR and lowering the adenoma miss rate. Furthermore, there is no reason to believe that the same methods, with the proper adjustments and training sets, could not work to tackle other real-time needs in endoscopy.

Supplementary Material

Note: To access the supplementary material accompanying this article, visit the online version of *Gastroenterology* at www.gastrojournal.org, and at https://doi.org/10.1053/j.gastro.2018.06.037.

References

1. American Cancer Society. Cancer facts and figures 2016. Available at: https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2016.html. Accessed September 3, 2018.
2. Strum WB. Colorectal adenomas. N Engl J Med 2016;374:1065–1075.
3. Russell LB, Kuntz KM, Lansdorp-Vogelaar, et al. A systematic comparison of microsimulation models of colorectal cancer: the role of assumptions about adenoma progression. Med Decis Mak 2011;31:530–539.
4. Winawer SJ, Zauber AG, Ho MN, et al. Prevention of colorectal cancer by colonoscopic polypectomy. N Engl J Med 1993;329:1977–1981.
5. Patel SG, Ahnen DJ. Prevention of interval colorectal cancers: what every clinician needs to know. Clin Gastroenterol Hepatol 2014;12:7–15.
6. Pohl H, Robertson DJ. Colorectal cancers detected after colonoscopy frequently result from missed lesions. Clin Gastroenterol Hepatol 2010;8:858–864.
7. Leufkens AM, van Oijen MGH, Vleggaar FP, et al. Factors influencing the miss rate of polyps in a back-to-back colonoscopy study. Endoscopy 2012;44:470–475.
8. Corley DA, Jensen CD, Marks AR, et al. Adenoma detection rate and risk of colorectal cancer and death. N Engl J Med 2014;370:1298–1306.
9. Than M, Witherspoon J, Shami J, et al. Diagnostic miss rate for colorectal cancer: an audit. Ann Gastroenterol 2014;28:94.
10. Anderson JC, Butterly LF. Colonoscopy: quality indicators. Clin Transl Gastroenterol 2015;6:e77.

11. **Kaminski MF, Wieszczy P**, Rupinski M, et al. increased rate of adenoma detection associates with reduced risk of colorectal cancer and death. Gastroenterology 2017; 153:98–105.
12. GI Quality Measures for 2017 Released in MACRA Final Rule. Available at: http://partner.gastro.org/gi-quality-measures-for-2017-released-in-macra-final-rule. Accessed September 3, 2018.
13. **Bond A, Sarkar S**. New technologies and techniques to improve adenoma detection in colonoscopy. World J Gastrointest Endosc 2015;7:969.
14. Hassan C, Senore C, Radaelli F, et al. Full-spectrum (FUSE) versus standard forward-viewing colonoscopy in an organised colorectal cancer screening programme. Gut 2017;66:1949–1955.
15. Waldmann E, Britto-Arias M, Gessl I, et al. Endoscopists with low adenoma detection rates benefit from high-definition endoscopy. Surg Endosc 2015;29:466–473.
16. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw 2015;61:85–117.
17. Baldi P, Chauvin Y. Neural networks for fingerprint recognition. Neural Comput 1993;3:402–418.
18. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv 2014;1409.1556.
19. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. IEEE Conf Comput Vis Pattern Recogn 2016:770–778.
20. Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection. Proc IEEE Conf Comput Vis Pattern Recogn 2016:779–788.
21. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. Proc IEEE Conf Comput Vis Pattern Recogn 2015:1–9.
22. Graves A, Mohamed AR, Hinton G. Speech recognition with deep recurrent neural networks. Acoustics Speech Signal Process 2013:6645–6649.
23. Wu Y, Schuster M, Chen Z, et al. Google's neural machine translation system: bridging the gap between human and machine translation. arXiv 2016;1609.08144.
24. Silver D, Huang A, Maddison C, et al. Mastering the game of Go with deep neural networks and tree search. Nature 2016;529:484–489.
25. Wu L, Baldi P. Learning to play GO using recursive neural networks. Neural Netw 2008;21:1392–1400.
26. Baldi P, Sadowski P, Whiteson D. Searching for exotic particles in high-energy physics with deep learning. Nat Commun 2014;5:4308.
27. Shimmin C, Sadowski P, Baldi P, et al. Decorrelated jet substructure tagging using adversarial neural network. Physical Rev D 2017;96:074034.
28. Fooshee D, Mood A, Gutman E, et al. deep learning for chemical reaction prediction. Mol Syst Des Eng 2017;10.1039/C7ME00107J.
29. Di Lena P, Nagata K, Baldi P. Deep architectures for protein contact map prediction. Bioinformatics 2012; 28:2449–2457.
30. Wang J, Fang Z, Lang N, et al. A multi-resolution approach for spinal metastasis detection using deep siamese neural networks. Comput Biol Med 2017; 84:137–146.
31. Wang J, Ding H, Azamian F, et al. Detecting cardiovascular disease from mammograms with deep learning. IEEE Trans Biomed Imaging 2017;36:1172–1181.
32. Chang P, Su L, Baldi P, et al. Deep learning convolutional neural networks accurately classify genetic mutations in gliomas. AJNR Am J Neuroradiol 2018;39:1201–1207.
33. Esteva A, Kuprel B, Novoa R, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115–118.
34. Baldi P. Deep learning in biomedical data science. Annu Rev Biomed Data Sci 2018;1:181–205.
35. **Russakovsky O, Deng J**, Su H, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis 2015; 115:211–252.
36. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv 2015;1502.03167.
37. Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. Proc IEEE Conf Comput Vis Pattern Recogn 2016:2818–2826.
38. Chollet F. *Keras*. San Francisco, CA: GitHub 2015.
39. Abadi M, Agarwal A, Barham P, et al. TensorFlow: Large-scale machine learning on heterogeneous systems. Available at: https://www.tensorflow.org/. Published 2015.
40. Hinton GE, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors. arXiv 2012;1207.0580.
41. Baldi P, Sadowski P. The dropout learning algorithm. Artif Intell 2014;210:78–122.
42. Chatfield K, Simonyan K, Vedaldi A, et al. Return of the devil in the details: delving deep into convolutional nets. arXiv 2014;1405.3531.
43. Bernal J, Sánchez J, Vilarino F. Towards automatic polyp detection with a polyp appearance model. Pattern Recogn 2012;45:3166–3182.
44. Ribeiro E, Uhl A, Häfner M. Colonic polyp classification with convolutional neural networks. IEEE 29th Int Symp Comput Based Med Syst 2016:253–258.
45. Soetikno R, Friedland S, Kaltenbach T, et al. Non-polypoid (flat and depressed) colorectal neoplasms. Gastroenterology 2006;130:566–576.
46. Park SY, Sargent D. Colonoscopic polyp detection using convolutional neural networks. SPIE Med Imaging 2016;978528–978528.
47. **Bernal J, Tajkbaksh N**, Sánchez F, et al. Comparative validation of polyp detection methods in video colonoscopy: results from the MICCAI 2015 Endoscopic Vision Challenge. IEEE Trans Med Imaging 2017;36:1231–1249.
48. Olympus Annual Report 3. Available at: https://www.olympus-global.com/ir/data/annualreport/pdf/ar2017e_A3.pdf. Published 2018.

Author names in bold designate shared co-first authorship.

Received September 28, 2017. Accepted June 11, 2018.

Reprint requests
Address requests for reprints to: Pierre Baldi, PhD, 4038 Bren Hall, University of California, Irvine, California 92697-3435. e-mail: pfbaldi@uci.edu.

Acknowledgments
We thank Anish Patel, Junhee Kim, Gregory Albers, and Kenneth Chang for their support on this project. We also acknowledge a hardware gift from NVDIA and thank Yuzo Kanomata for computing support.

Author contributions
PB and WK conceived the concept of this project. WK managed the data acquisition, data labeling, and video study and contributed to the analysis of the results and manuscript drafting and editing. TA and MM contributed to labeling the data (TA significantly). PB and GU developed the technology. GU implemented the machine-learning architectures, ran the main experiments, and contributed to the analysis of the results and to manuscript drafting and editing. PT, MM, FJ, and WK served as expert reviewers in the video study. PB designed the algorithms and methodology, oversaw the experiments, analyzed the results, and contributed to manuscript drafting and editing. All authors approved the manuscript.

Conflicts of interest
Authors have no financial, professional, or personal conflict of interest.

Funding
Supported in part by grants NIH GM123558 and NSF IIS-1550705 to PB.

Supplementary Material

Architectures, Training, and Evaluation

Neural Network Architectures

We performed and evaluated experiments separately for polyp detection and polyp localization for clarity and to exclude confounding factors. In a use-case scenario, these can be performed by a single CNN model simultaneously. Polyp detection was implemented as a binary classification task of images by predicting the probability of whether an image would contain at least 1 polyp.

For localizing polyps, we used an approach that leaves the model's run time nearly unchanged to perform localization in real time. The neural network was trained to directly predict the center coordinates and the height and width of a bounding box that covers the polyp(s) in an image. An efficient approach to handling multiple polyp appearances in a single image is to predict not just 1 but several bounding boxes at once. This technique is a central aspect of "you only look once" (YOLO). In YOLO, the network has outputs spaced over a regular grid of the image (eg, 8 × 8), and each output predicts a confidence that an object of interest is within its grid cell and the object's location and size (ie, 5 values are predicted). This allows the detection of multiple polyps simultaneously and incurs only a negligible computational overhead compared with a model that can predict the location of only a single object. We experimented with a variation of YOLO for better localizing single objects in an image, which we named the "internal ensemble." For this all fully connected layers are replaced with convolutional layers. The last convolutional layer has 5 filters maps and will intrinsically have its outputs spaced over a grid over the input image (eg, 7 × 7). Each grid cell predicts its confidence (sigmoid unit), the position of the polyp relative to the grid cell center, and its size. The overall model output is the weighted sum of all 49 (7 × 7) adjusted position and size predictions, weighted with the predicted confidences. Hence, the model learns to use the confidence predictions (and thus weights) in an optimal way.

Supplementary Table 1 presents the CNNs that were trained for polyp detection and localization. The convolutional layers of PI-CNN1–3 in Supplementary Table 1 are pre-initialized with weights that were obtained by training a closely related model on the ImageNet challenge dataset[35]—these model architectures are based on the VGG16,[18] VGG19,[18] and ResNet50[19] models, respectively. We retrained all layers and weights of these CNNs on our polyp dataset. Networks NPI-CNN1 and -2 are not pre-initialized this way, but rather with random weights and thus trained from "scratch." However, they are smaller to decrease overfitting and thus faster than the PI models. The weights of all layers of NPI models and those of all fully connected layers of PI models were initialized with an adaptively scaled normal distribution.[S1]

We experimented with a variety of architectures for non–pre-trained CNNs other than the ones presented in Supplementary Table 1. Although the reported networks had the lowest validation-set classification error rates, it should be noted that many alternative architectures reached very similar validation- and test-set accuracies. The number of layers, filters per layer, and maximum pooling and subsampling operations was chosen to keep the computational cost of evaluating the model low and to never pool by more than a factor of 2 between convolutional operations because this would negatively affect detection accuracy when using small filters.

The following list provides overview of the different layer types commonly used in deep learning:

- Convolutional layers convolve their input feature maps with a set of trainable filters to extract features. A sequence of convolutional layers extracts a hierarchy of features from implicitly increasingly large regions of the input image, with larger regions for layers higher in the hierarchy (further away from the input). The filter weights of a given layer are shared and used across the entire input field, which implies that convolutional layers are implicitly heavily regularized and usually contribute a relatively small amount of trainable weights compared with fully connected layers. Convolutions are at the core of most computer-vision and signal-processing models.
- Fully connected layers correspond to a matrix–matrix multiplication with the output of the previous layer, where every entry in the matrix is a trainable parameter.
- Batch normalization layers are introduced to prevent the distribution of each layer's outputs from changing during training and have been shown to speed up learning and improve classification performance in some cases.[36] They normalize feature maps such that each resulting feature has 0 mean and unit standard deviation. The mean and standard deviation for the z-score normalization are estimated with a moving average during training.
- Nonlinear activation functions are applied to the output of virtually every layer in the network. Without these, most deep network architectures would be mathematically equivalent to a simpler shallow linear model. Note that operations such as convolution, batch normalization, and matrix–matrix multiplication are intrinsically linear.
- Maximum pooling layers summarize the outputs of neighboring regions in the same feature map. The output of a maximum pooling unit corresponds to the maximum of its input features taken over a certain region. In this study, the output was generated for every other location (ie, with stride 2) by considering the 2 × 2 neighborhood region of the location, denoted as p in Supplementary Table 1.

Training

When training a machine learning model to perform regression, one typically optimizes the L2 loss because it is the most natural loss function for general regression tasks

for parameters such as polyp size and location. However, in this case, it implies treating the location coordinates and bounding-box dimensions independently from each other. This might not necessarily be optimal because highly accurate bounding-box width and height predictions are useless if the location predictions are bad. Thus, we also experimented with optimizing the Dice coefficient directly using a custom loss function, because it is a true measure of segmentation performance.

All weights of polyp localization models were initialized with weights of the corresponding trained polyp-detection CNN to accelerate training and facilitate convergence. The neural networks were trained with the ADAM optimizer.[S2] We used a batch size of 16 and optimized the initial learning rate for each architecture individually. We decreased the learning rate automatically when the validation accuracy did not improve for 10 consecutive epochs and stopped training after it had not improved for 50 epochs. In our experiments, we evaluated model performance by applying a 7-fold cross-validation procedure, because this decreases the variance of performance estimates compared with a simple single split. In each cross-validation run, the dataset was split into a different training, development, and test subset. The union of all test sections of the cross-validation splits recovers the original dataset (ie, the test sections are mutually exclusive). The CNN was trained on the training segment, whereas the development segment was used to monitor the CNN's accuracy during training for early stopping and hyperparameter optimization. All reported results were obtained on the test segments, which was left untouched during the training process. Furthermore, the weights of all CNNs were independently retrained from their initial configuration for each data split.

Performance Evaluation and Metrics

For evaluation of our (binary) classification results, we relied mainly on 2 metrics: accuracy and AUC. Machine learning models for classification typically predict a value that corresponds to a class probability. For a given threshold, performance is determined by the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Other quantities of interest such as accuracy, true positive rate (TPR; also called sensitivity or recall), and FPR can be easily obtained (see Baldi and Brunak[S3] for more details):

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$TPR = Sensitivity = \frac{TP}{P} = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN}$$

In contrast, the AUC is invariant to the choice of threshold; it is the area under the receiver operating characteristic curve. The receiver operating characteristic is obtained by plotting the TPR against the FPR for all thresholds in the range [0, 1]. Thus, the best possible AUC is 1.

To evaluate polyp localization and segmentation performance, we used the commonly used Sørensen–Dice coefficient D (also known as F1 score), which compares the relative area of overlap between the predicted bounding box (A) and the ground-truth (B) bounding box:

$$D(A, B) = 2\frac{|A \cap B|}{|A| + |B|}$$

The Dice coefficient is equivalent to Jaccard index J and also is known as the "intersection over union" (IoU) or Tanimoto distance. They can be trivially computed from each other using these relations:

$$D = \frac{2J}{1 + J} \quad J = IoU = \frac{D}{2 - D}$$

Polyp Detection

Supplementary Figure 2 shows the receiver operator characteristic curve for the 2 NPI models and the 3 PI CNNs when trained and evaluated on the 8,641 colonoscopy images (shown are test-set predictions for a single cross-validation split of the 5 models). The clinically most relevant section is the one with low FPRs (ie, high specificity), because excessive false alerts would be distracting and eventually be ignored by the colonoscopist. The 3 PI models are clearly better than the NPI models in this regime.

Colonoscopy Video Study With Expert Colonoscopists

The following tables present a detailed breakdown of results of expert review of videos with and without CNN assistance compared with CNN alone. The tables list polyp characteristics (size, shape, and location). Supplementary Table 2 also presents the confidence of the expert reviewer on the presence of the polyp and the number of polyps in 8 contiguous frames predicted by the CNN as "positive" (all >50% confidence to contain a polyp).

Video Study 1

See Supplementary Tables 2 and 3.

Video Study 2

Supplementary Table 4 presents an expert review of 11 deliberately challenging videos (withdrawal of colonoscope without closing in on already seen polyps).

Statistics of Number of Polyps Per Video

See Supplementary Table 5.

Effect of Post-Processing on Sensitivity and Specificity in Expert Video Study

The CNN predictions on videos were filtered by a window size of 8 consecutive frames to remove likely false-positive findings. Supplementary Figure 1 shows the effect of different settings of this window size on the sensitivity and specificity of CNN predictions at a fixed threshold (on data from the first expert video study). A CNN prediction for any given frame is counted as positive if it is part of at least 1 window of consecutive predictions greater than 40% predicted confidence for polyp presence; otherwise, it is counted as negative. We found that a window size of 8 results in a well-balanced ratio of sensitivity to specificity.

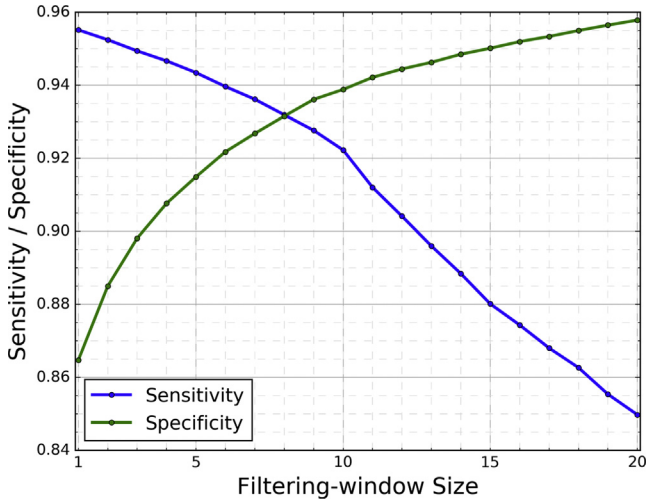
Supplementary References

S1. He K, Zhang X, Ren S, et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. Proc IEEE Int Conf Computer Vision 2015:1026–1034.

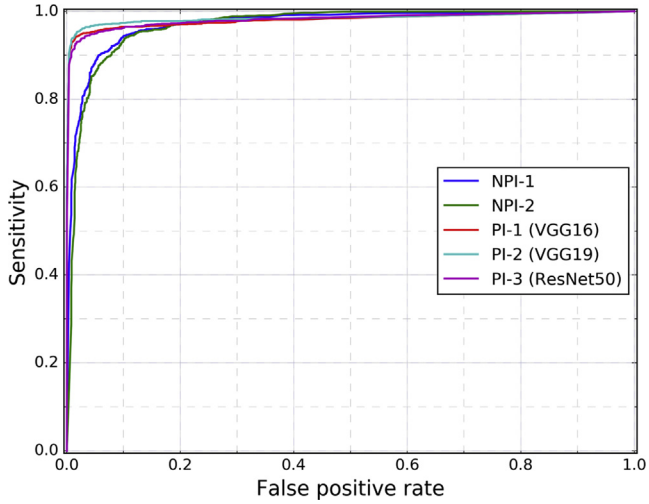
S2. **Kingma DP, Ba JL**, Adam: A Method for Stochastic Optimization. Proceedings of the 3rd International Conference on Learning Representations (ICLR) 2015.

S3. Baldi P, Brunak S. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 2000:412–424.

Author names in bold designate shared co-first authorship.



Supplementary Figure 1. Effect of the size of the filtering window on the sensitivity and specificity of CNN predictions using a fixed threshold of 0.4.



Supplementary Figure 2. Receiver operator characteristic curve for all 5 CNN architectures trained on subsets of the 8641 colonoscopy images. Results obtained on the test splits of the 8641 colonoscopy images.

Supplementary Table 1. Summary of Architectures of Deep CNNs Used in This Study

Model	Initial weights	Architecture
NPI-CNN1	—	64c-p-64c-p-128c-p-128c-p-256c-p-245c-p-512fc-512fc
NPI-CNN2	—	24c-bn-p-24c-bn-p-24c-bn-p-Incept(16)-bn-Incept(16)-64c-p-64c-768fc-768fc
PI-CNN1	VGG16	64c-64c-p-128c-128c-p-$(256c)^3$-p-$(512c)^3$-p-$(512c)^3$-p-1560fc-1560fc
PI-CNN2	VGG19	64c-64c-p-128c-128c-p-$(256c)^4$-p-$(512c)^4$-p-$(512c)^4$-p-1560fc-1560fc
PI-CNN3	ResNet50	64c-bn-p-Cb-$(Ib)^2$-Cb-$(Ib)^3$-Cb-$(Ib)^5$-Cb-$(Ib)^2$-p-1024fc

NOTE. The first layer is the leftmost entry and the last hidden layer is the rightmost entry per row. Exponentiation means repetition (eg, $[256c]^3$ corresponds to 3 convolutional layers with 256 filters each). Ib and Cb are structures containing 6 and 8 layers, respectively (for details, see Grave et al[22]). Convolutional filter sizes are typically 3×3.
bn, batch-normalization layer; c, convolutional layer; fc, fully connected layer; Incept, inception module[30]; p, maximum pooling layer.

Supplementary Table 2. List of All Unique Polyps Found in the First Set of 9 Videos, With Their Location and Estimated Size

Expert polyp encounter				CNN alone (8 contiguous frames)			Polyp characteristics			
Without CNN	With CNN	Removed or missed	Polyp ID	CNN^+	CNN^-	CNN^+, %	Size (*mm*)	Shape	Location	Confidence
No	No	NA	NA	2708	28,146	9	NA	NA	NA	NA
No	Yes	Missed[a]	VID-1-3	6	0	100	2	Sessile	Hepatic flexure	Low
			VID-1-5	5	0	100	4	Sessile	Transverse	High
			VID-3-3	5	0	100	2	Sessile	Rectum	High
			VID-4-4	17	0	100	7	Sessile	Ascending	High
			VID-5-2	6	0	100	2	Sessile	Descending	Low
			VID-6-3	3	0	100	2	Sessile	Descending	Low
			VID-7-2	12	0	100	6	Sessile	Descending	Low
			VID-7-6	8	0	100	6	Sessile	Descending	Low
			VID-8-2	1	0	100	2	Sessile	Cecum	Low
			Mean	6.8	0.0	100	3.7			
Yes	Yes	Not removed	VID-1-2	21	1	97	2	Sessile	Hepatic flexure	High
			VID-1-4	7	1	94	6	Sessile	Ascending	High
			VID-2-1	65	5	93	2	Sessile	Sigmoid	High
			VID-3-4	6	0	100	1	Sessile	Rectum	Low
			VID-4-2	128	0	100	3	Sessile	Ascending	High
			VID-6-2	2	6	26	1	Sessile	Ascending	High
			VID-7-1	24	0	100	3	Sessile	Ascending	High
			VID-7-5	164	0	100	3	Sessile	Sigmoid	High
			Mean	52.1	1.5	88.6	2.6			
Yes	Yes	Removed	VID-1-1	149	78	66	6	Sessile	Hepatic flexure	High
			VID-1-6	90	23	80	10	Flat	Ascending	High
			VID-1-7	86	6	94	3	Sessile	Descending	High
			VID-1-8	200	35	85	2	Sessile	Rectum	High
			VID-3-1	24	4	86	2	Sessile	Sigmoid	High
			VID-3-2	151	7	95	2	Sessile	Sigmoid	High
			VID-3-5	201	0	100	2	Sessile	Sigmoid	High
			VID-4-1	624	15	98	15	Flat	Ascending	High
			VID-4-3	102	4	96	1	Sessile	Ascending	High
			VID-5-1	256	4	98	6	Sessile	Descending	High
			VID-6-1	2,188	280	89	10	Sessile	Ascending	High
			VID-7-3	238	29	89	5	Sessile	Descending	High
			VID-7-4	290	17	94	5	Sessile	Sigmoid	High
			VID-7-7	675	63	92	5	Sessile	Sigmoid	High
			VID-8-1	85	10	89	1	Sessile	Cecum	High
			VID-8-3	139	17	89	2	Sessile	Descending	High
			VID-8-4	126	6	95	3	Sessile	Descending	High
			VID-9-1	221	76	74	6	Sessile	Descending	High
			VID-9-10	86	8	92	4	Sessile	Descending	High
			VID-9-11	100	3	97	20	Sessile	Sigmoid	High
			VID-9-2	196	7	97	2	Sessile	Descending	High
			VID-9-3	53	0	100	4	Sessile	Descending	High
			VID-9-4	24	0	100	4	Sessile	Descending	High
			VID-9-5	67	0	100	4	Sessile	Cecum	High
			VID-9-6	51	3	95	4	Sessile	Cecum	High
			VID-9-7	33	0	100	2	Sessile	Cecum	High
			VID-9-8	124	3	98	3	Sessile	Cecum	High
			VID-9-9	92	0	100	4	Sessile	Descending	High
			Mean	238.2	24.9	92.4	4.9			

[a]Missed polyps are those that were found with CNN assistance but not by expert review of videos without the CNN overlay.

Supplementary Table 3. Contingency Analysis Assessing Performance of CNN and CNN-Assisted Expert Review of the 9 Colonoscopy Videos

	CNN alone			
Expert + CNN	Polyp	Non-polyp		
Polyp	48,911	3,577	Sensitivity	0.93
Non-polyp	17,407	236,728	Specificity	0.93
			PPV	0.74
			NPV	0.985

NOTE. Analysis is based on comparing expert annotations (on CNN-overlaid videos) and CNN predictions on a "frame-by-frame" basis across videos and assumes CNN-assisted expert review is the "gold standard."
NPV, negative predictive value; PPV, positive predictive value.

Supplementary Table 4. List of All Unique Polyps Found in Second Set of 11 Videos and Their Location and Estimated Size

Polyp ID	Size (*mm*)	Shape	Removed?	Location
10-1	3	Sessile	Yes	Cecum
10-2	2	Sessile	Yes	Cecum
10-3	4	Flat	Yes	Cecum
10-4	3	Sessile	Yes	Sigmoid
10-5	5	Sessile	Yes	Rectum
10-6	4	Flat	Yes	Rectum
11-1	2	Sessile	Yes	Cecum
11-2	2	Sessile	Yes	Cecum
11-3	2	Sessile	Yes	Cecum
11-4	5	Sessile	Yes	Ascending
11-5	3	Sessile	Yes	Ascending
11-6	2	Sessile	Yes	Ascending
11-7	2	Sessile	Yes	Ascending
11-8	3	Sessile	Yes	Ascending
11-9	1	Sessile	No	Ascending
11-10	1	Sessile	No	Ascending
11-11	3	Sessile	Yes	Ascending
11-12	3	Sessile	Yes	Ascending
11-13	3	Sessile	Yes	Ascending
11-14	4	Sessile	Yes	Ascending
11-15	2	Sessile	No	Ascending
11-16	2	Sessile	No	Ascending
11-17	3	Sessile	Yes	Ascending
11-18	2	Sessile	Yes	Ascending
11-19	2	Sessile	Yes	Ascending
11-20	3	Sessile	Yes	Ascending
11-21	2	Sessile	Yes	Ascending
11-22	3	Sessile	Yes	Ascending
11-23	3	Sessile	Yes	Hepatic flexure
11-24	2	Sessile	Yes	Hepatic flexure
11-25	3	Sessile	Yes	Hepatic flexure
11-26	4	Sessile	Yes	Hepatic flexure
11-27	1	Sessile	No	Hepatic flexure
11-28	2	Sessile	No	Transverse
11-29	2	Sessile	No	Transverse
11-30	1	Sessile	No	Splenic flexure
11-31	2	Sessile	No	Splenic flexure
11-32	7	Pedunculated	Yes	Descending
12-1	2	Sessile	Yes	Cecum
12-2	1	Sessile	Yes	Ascending
15-1	2	Sessile	No	Ascending
16-1	1	Sessile	No	Cecum
16-2	2	Sessile	Yes	Hepatic flexure
16-3	4	Flat	Yes	Sigmoid
16-4	1	Sessile	No	Sigmoid
17-1	2	Sessile	No	Descending
17-2	4	Sessile	Yes	Descending
18-1	2	Sessile	No	Ascending
18-2	7	Sessile	Yes	Hepatic flexure
18-3	1	Sessile	Yes	Splenic flexure
19-1	3	Sessile	No	Descending
20-1	2	Sessile	Yes	Cecum
20-2	1	Sessile	Yes	Hepatic flexure
20-3	2	Sessile	Yes	Transverse
20-4	4	Sessile	Yes	Transverse
20-5	6	Sessile	Yes	Transverse
20-6	4	Sessile	Yes	Splenic flexure
20-7	2	Sessile	Yes	Splenic flexure
20-8	3	Sessile	Yes	Splenic flexure
20-9	3	Sessile	Yes	Descending
20-10	3	Sessile	Yes	Descending
20-11	2	Sessile	Yes	Descending
20-12	5	Sessile	Yes	Descending
20-13	3	Sessile	Yes	Descending
20-14	3	Sessile	Yes	Descending
20-15	4	Sessile	Yes	Descending
20-16	3	Sessile	Yes	Descending
20-17	2	Sessile	Yes	Descending
20-18	4	Sessile	Yes	Sigmoid
20-19	3	Sessile	Yes	Sigmoid
20-20	4	Sessile	Yes	Sigmoid
20-21	4	Sessile	Yes	Sigmoid
20-22	2	Sessile	Yes	Sigmoid

Supplementary Table 5. Total Number of Polyps Found per Video in 20 Videos

Video ID	Polyps, n
1	8
2	1
3	5
4	4
5	2
6	3
7	7
8	4
9	11
10	6
11	32
12	2
13	0
14	0
15	1
16	4
17	2
18	3
19	1
20	22