Contents lists available at ScienceDirect

Medical Image Analysis

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

Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps

Kristoffer Wickstrøm[1,*], Michael Kampffmeyer[1], Robert Jenssen[1]

Department of Physics and Technology, UiT The Arctic University of Norway, Tromsø NO-9037, Norway

ARTICLE INFO

Article history:
Received 10 May 2019
Revised 14 November 2019
Accepted 14 November 2019
Available online 20 November 2019

Keywords:
Polyp segmentation
Decision support systems
Fully convolutional networks
Monte carlo dropout
Guided backpropagation
Monte carlo guided backpropagation

ABSTRACT

Colorectal polyps are known to be potential precursors to colorectal cancer, which is one of the leading causes of cancer-related deaths on a global scale. Early detection and prevention of colorectal cancer is primarily enabled through manual screenings, where the intestines of a patient is visually examined. Such a procedure can be challenging and exhausting for the person performing the screening. This has resulted in numerous studies on designing automatic systems aimed at supporting physicians during the examination. Recently, such automatic systems have seen a significant improvement as a result of an increasing amount of publicly available colorectal imagery and advances in deep learning research for object image recognition. Specifically, decision support systems based on Convolutional Neural Networks (CNNs) have demonstrated state-of-the-art performance on both detection and segmentation of colorectal polyps. However, CNN-based models need to not only be precise in order to be helpful in a medical context. In addition, interpretability and uncertainty in predictions must be well understood. In this paper, we develop and evaluate recent advances in uncertainty estimation and model interpretability in the context of semantic segmentation of polyps from colonoscopy images. Furthermore, we propose a novel method for estimating the uncertainty associated with important features in the input and demonstrate how interpretability and uncertainty can be modeled in DSSs for semantic segmentation of colorectal polyps. Results indicate that deep models are utilizing the shape and edge information of polyps to make their prediction. Moreover, inaccurate predictions show a higher degree of uncertainty compared to precise predictions.

1. Introduction

Colorectal Cancer (CRC) is one of the leading causes of cancer-related deaths worldwide (Siegel et al., 2017; Chen et al., 2016; Larsen, 2016), with an estimated five-year survival rate for an advanced stage CRC diagnosis of 14%. The estimated survival rate for early diagnosis is 90% (Larsen, 2016). Currently, the gold standard for CRC prevention is through regular colonoscopy screenings. One of the main tasks during a screening is to locate small abnormal growths called polyps, which are known to be possible precursors to CRC. Hence, increasing the detection rate of polyps is an important component for reducing mortality rates. However, such screenings are manual procedures performed by physicians and are therefore affected by human factors such as fatigue and experience. One study has estimated the polyp miss rate during a screening to be between 8–37%, depending on the size and type of the polyps (Van Rijn et al., 2006). A possible method for increasing polyp detection rate is to design Decision Support Systems (DSSs), which could aid physicians during or after the procedure. A dependable and robust DSS would have the advantage of not being influenced by human factors and could also provide a second opinion for inexperienced practitioners.

One popular approach for developing DSSs has been through machine learning, with promising results on a range of different tasks like brain tumor segmentation (Havaei et al., 2017), retinal vessel segmentation (Guo et al., 2019), melanoma lesion segmentation (Nida et al., 2019), and colorectal polyp detection (Bernal et al., 2015; 2014; Liu, 2017; Ribeiro et al., 2016). In the context of CRC prevention, there have been a number of studies on the detection of polyps with encouraging results (Tajbakhsh et al., 2016; Hwang et al., 2007; Alexandre et al., 2007; Wimmer et al., 2016; Häfner et al., 2015), but polyp segmentation has proven to be a challenging task and the necessary precision has been difficult to obtain (Bernal et al., 2015; 2014; Condessa and Bioucas-Dias, 2012).

[*] Corresponding author.
E-mail address: kristoffer.k.wickstrom@uit.no (K. Wickstrøm).
[1] UiT Machine Learning Group (http://machine-learning.uit.no).

https://doi.org/10.1016/j.media.2019.101619

However, as a consequence of increasing amounts of publicly available colon imagery combined with advances in deep learning research for image analysis, recent studies based on deep learning for colorectal polyp segmentation have shown promising results and a significant increase in precision (Vázquez et al., 2016; Brandao et al., 2017; Urban et al., 2018).

High precision is a crucial component of any reliable DSS, but other constituents are also vital in order to engineer dependable DSSs. Physicians are tasked with making decisions that can have fatal consequences and they go to great lengths in order to ensure that the decision they make is likely to have a favorable outcome. Therefore, a trustworthy DSS should provide a measure of uncertainty to accompany its prediction such that physicians can make well-informed decisions. Another integral part of a dependable DSS is to communicate to the user what factors influences a prediction. Without such information, the user can not determine if the model is detecting features that are actually associated with the disease in question or if it is exploiting artifacts in the data. For instance, a study by Zech et al. (2018) uncovered that a deep learning model tasked with diagnosing disease from x-ray images had learned to exploit information in metal tokens included in the x-ray images for inference instead of detecting disease-specifics features. When the model is then presented with an image without these artifacts the precision drops considerably.

Despite the obvious benefit of increased performance, systems based on deep learning have no inherent way of representing the uncertainty associated with a model's prediction nor do they provide any indication as to what features in the input influences a particular prediction. This lack of theoretical understanding for the underlying mechanics of deep models have resulted in deep learning based models often being referred to as ”black boxes” (Alain and Bengio, 2017; Shwartz-Ziv and Tishby, 2017; Yu and Príncipe, 2018). Multiple recent studies have proposed methods that, to some extent, address the lack of transparency (Gal and Ghahramani, 2016; Kendall and Gal, 2017; Springenberg et al., 2015; Zeiler and Fergus, 2014; Bach et al., 2015; Simonyan et al., 2013), and they have seen some use in analysis of medical images (Dubost et al., 2019; Zech et al., 2018) However, these methods have yet to be utilized in DSSs for colorectal polyp segmentation based on deep learning.

Our contributions are the following:[2]

- We incorporate and develop recent advances in the field of deep learning for semantic segmentation of colorectal polyps in order to create deep models that provide uncertainty measures along with their prediction. Results indicate that erroneous predictions show a significantly higher degree of uncertainty compared to correct predictions. Furthermore, we model input feature importance to create interpretable deep models. Results show that our models are considering shape and edge information in order to segment polyps.
- We propose a novel method for estimating uncertainty in the importance of input features, which we refer to as Monte Carlo Guided Backpropagation, and demonstrate how this method can be used in the context of colorectal polyp segmentation.

To the authors' knowledge, none of the above points have been previously explored in the context of semantic segmentation of colorectal polyps.

2. Models and methods

In this section we introduce Fully Convolutional Networks (FCNs) and describe the three architectures utilized in this study. Next, we explain how we incorporate uncertainty and interpretability in deep learning based DSSs (Sections 2.2 and 2.3). Finally, we present our method for estimating the uncertainty associated with the importance of input features (Section 2.4).

2.1. Fully convolutional networks

FCNs are CNNs particularly suited to tackle per pixel prediction problems like semantic segmentation, i.e. providing a probability score for what class each pixel belongs to. For instance, in the case of semantic segmentation of colorectal polyps, each pixel is labeled as a polyp or as part of the colon (background class). Segmentation is considered a more challenging task than detecting or localizing an object in an image, but provides more information. The shape information provided by a meaningful segmentation map can for example be used to study anatomical structures or inspect other regions of interest (Sharma et al., 2010).

We investigate three architectures for the task of polyp segmentation, namely the Fully Convolutional Network 8 (FCN-8) (Shelhamer et al., 2017), U-Net (Ronneberger et al., 2015) and SegNet (Badrinarayanan et al., 2017) for the following reasons. These networks have been applied in a number of different domains and are chosen to form a well-understood foundation for our studies. This enables uncertainty and interpretability experiments to be the main focus. Previous use of the FCN-8 for polyp segmentation has shown promising results (Vázquez et al., 2016; Brandao et al., 2017). SegNet has been shown to achieve comparable results to the FCN-8 in some applications but is a less memory intensive approach with fewer parameters to optimize. U-Net has previously demonstrated encouraging results on medical tasks and does also contain fewer parameters than the FCN-8, thus providing a lightweight alternative. We include these different networks in this study in order to compare what features are considered important by different models and how uncertainty estimates differ among networks. The interested reader can find a detailed description along with figures of the three models in Appendix A.

2.2. Uncertainty in fully convolutional networks

Despite their success on a number of different tasks, CNNs are not without flaws. One of these flaws, which becomes especially apparent for medical applications, is their inability to provide any notion of uncertainty in their prediction. When a physician is considering the symptoms of a patient and contemplates what medication to prescribe there might be several viable options, and the final decision might spell the difference between a fatal or favorable outcome. Since the stakes are so high, physicians will have to weight the different options and reflect on which choice is most likely to have a favorable outcome. If a physician decides to consult a DSS based on a CNN, she or he would be presented with a recommendation that has no indication as to how likely a desirable outcome is, thus making it difficult for the physician to trust the system. Although the softmax output regularly found at the end of a CNN is sometimes interpreted as model confidence, this is generally ill-advised (Gal and Ghahramani, 2016) and other approaches must be considered.

In contrast, Bayesian models provide a framework which naturally includes uncertainty by modeling posterior distribution for the quantities in question. Given a dataset $\mathcal{D} \equiv \left\{\mathbf{x}_n \in \mathbb{R}^D, \mathbf{y}_n \in \mathbb{R}^C\right\}_{n=1}^N$, where $\mathbf{x}_n$ denotes an input vector and $\mathbf{y}_n$ denotes its corresponding one-hot encoded label vector, the predictive distribution of a Bayesian neural network for a new pair of

[2] This work significantly extends our preliminary study (Wickstrøm et al., 2018) by: (1) Including U-Net in our analysis; (2) significantly extending our experimental section by including new experiments on the 2015 MICCAI polyp detection challenge (Bernal et al., 2017) and the Endoscene dataset (Vázquez et al., 2016) (3) proposing a novel method for estimating uncertainty in the importance of input features and evaluating our proposed method on two polyp segmentation datasets; (4) providing a more thorough literature background discussion and placing our work into a broader context.

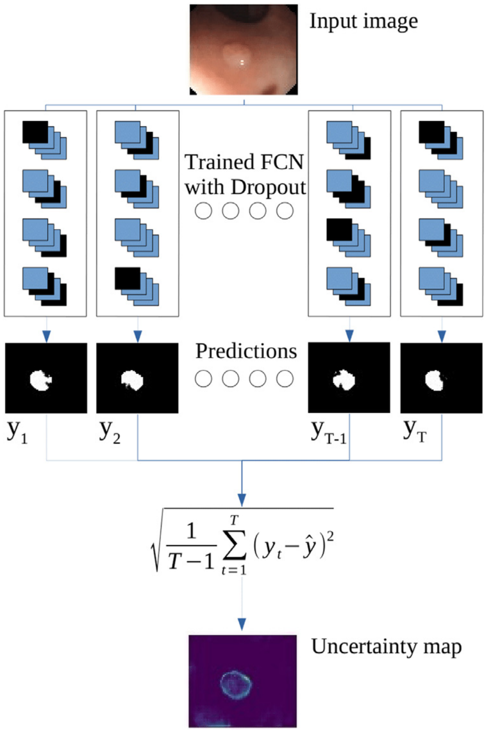


Fig. 1. Illustration of the Monte Carlo Dropout procedure. The same input image is passed through a trained FCN with Dropout applied T times, resulting in T different predictions. The standard deviation of each pixel is then estimated based on these T predictions.

samples $\{\mathbf{x}_*, \mathbf{y}_*\}$ can be modeled as:

$$p(\mathbf{y}_*|\mathbf{x}_*, \mathcal{D}) = \int p(\mathbf{y}_*|\mathbf{x}_*, \mathbf{W})p(\mathbf{W}|\mathbf{x}_*, \mathcal{D})d\mathbf{W} \qquad (1)$$

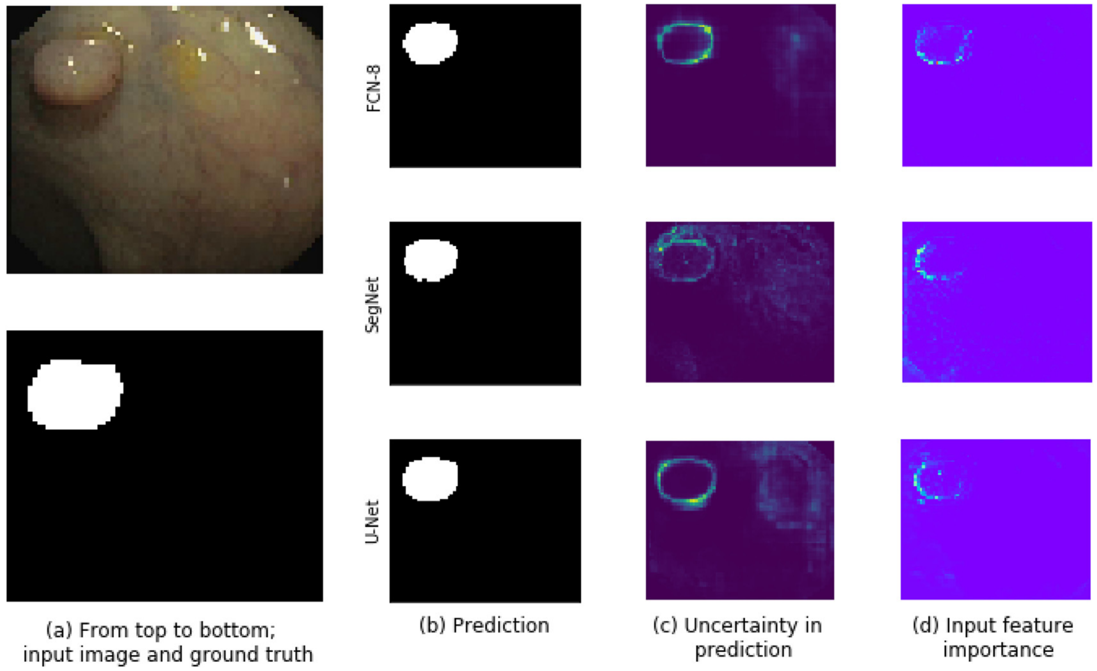
In Eq. (1), **W** refers to the weights of the model, $p(\mathbf{y}_*|\mathbf{x}_*, \mathbf{W})$ is the softmax function applied to the output of the model, denoted by $f_{\mathbf{W}}(\mathbf{x}_*)$, and $p(\mathbf{W}|\mathbf{x}_*, \mathcal{D})$ is the posterior over the weights which capture the set of plausible model parameters for the given data. Obtaining $p(\mathbf{y}_*|\mathbf{x}_*, \mathbf{W})$ only requires a forward pass of the network, but the inability to evaluate the posterior of the weights analytically makes Bayesian neural networks computationally infeasible. To sidestep the problematic posterior of the weights, (Gal and Ghahramani, 2016) proposed to incorporate Dropout as a method for sampling sets of weights from the trained network to approximate the posterior of the weights. The predictive distribution from Eq. (1) can then be approximated using Monte Carlo integration as follows:

$$p(\mathbf{y}_*|\mathbf{x}_*, \mathcal{D}) \approx \frac{1}{T}\sum_{t=1}^{T}\text{Softmax}(f_{\mathbf{W}_t^*}(\mathbf{x}_*)) \qquad (2)$$

where T is the number of sampled sets of weights and $\mathbf{W}^*{}_t$is a set of sampled weights. In practice, the predictive distribution from Eq. (2) can be estimated by running T forward passes of a model with Dropout applied to produce T predictions and then computing the standard deviation over the softmax outputs of the T samples. We will refer to these uncertainty estimates as uncertainty maps. This method of utilizing Dropout for sampling from the posterior of the predictive distribution is referred to as Monte Carlo Dropout, and the method is illustrated in Fig. 1.

2.3. Interpretability in fully convolutional networks

Another desirable property which CNNs lack is interpretability, i.e. being able to determine what features induce the network to produce a particular prediction. For instance, a physician might be interested in discerning what information the prediction of a given DSS is based on, and if it concurs with medical knowledge. A CNN-based DSS has no inherent way of providing such an explanation. However, several recent works have proposed different methods to increase network interpretability (Zeiler and Fergus, 2014; Bach et al., 2015). In this paper, we evaluate and develop the Guided Backpropagation (Springenberg et al., 2015) technique for FCNs on the task of semantic segmentation of colorectal polyps in order to



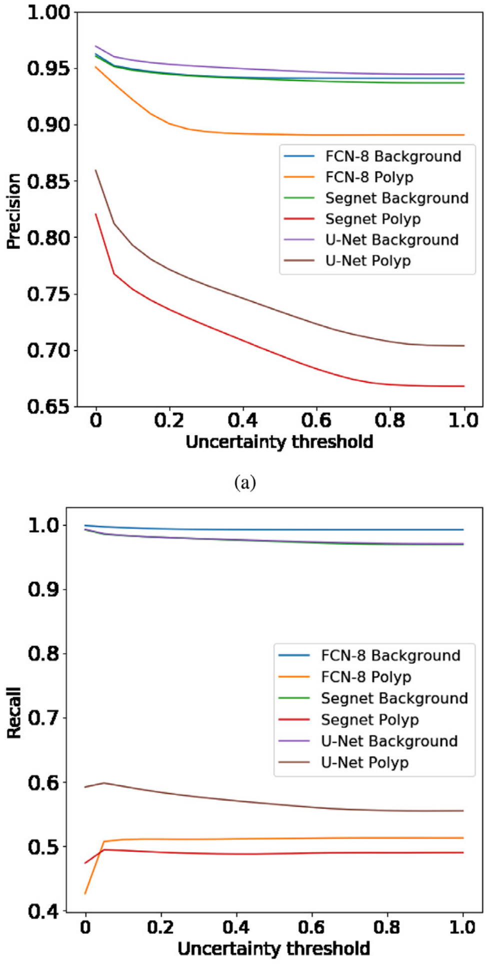
(a) From top to bottom; input image and ground truth

(b) Prediction

(c) Uncertainty in prediction

(d) Input feature importance

Fig. 2. Figure displays the prediction, uncertainty map, and interpretability map for the FCN-8, SegNet and U-Net, for the input image shown in the leftmost column. Best viewed in color.



(a)

(b)

Fig. 3. Precision and recall vs uncertainty plot for background and polyp class on the Endoscene test set.

assess which pixels in the input image the network deems important for identifying polyps. We choose Guided Backpropagation as it is known to produce clearer visualizations of salient input pixels compared to other methods (Zeiler and Fergus, 2014; Simonyan et al., 2013). We refer to these visualizations of salient pixels as interpretability maps.

The central idea of Guided Backpropagation is the interpretation of the gradients of the network with respect to an input image. Simonyan et al. (2013) exploited that, for a given image, the magnitude of the gradients indicate which pixels in the input image need to be changed the least to affect the prediction the most. By utilizing backpropagation (Rumelhart et al., 1988; Werbos, 1974), they obtained the gradients corresponding to each pixel in the input such that they could visualize what features the network considers essential. Springenberg et al. (2015) argued that positive gradients with a large magnitude indicate pixels of high importance while negative gradients with a large magnitude indicate pixels which the networks want to suppress. If these negative gradients are included in the visualization of important pixels it might result in noisy visualization of descriptive features. In order to avoid noisy visualizations the Guided Backpropagation procedure alters the backward pass of a neural network such that negative gradients are set to zero in each layer, thus allowing only positive gradients to flow backward through the network and highlighting pixels that the system finds important.

2.4. Monte carlo guided backpropagation: Uncertainty in input feature importance

To determine the uncertainty associated with an input feature's importance for the prediction, we propose a novel approach inspired by Monte Carlo Dropout combined with Guided Backpropagation. In Section 2.2 we discussed CNNs inability to produce any notion of uncertainty and described Monte Carlo Dropout, which provides a method to obtain approximate measures of uncertainty for CNNs by utilizing Dropout during inference. Accompanying a model's prediction with an uncertainty estimate adds the option to assess if a particular prediction is highly certain or a case that could require further analysis from a human expert. In Section 2.3 we described Guided Backpropagation, a technique developed to visualize the relative importance of input features for CNNs by considering the positive gradients from a backward pass through the network. But, determining the importance of the input features based on gradients from a single backward pass encounters the same issue we discussed regarding decisions based on predictions from a single forward pass. How confident are we that these features are important for the decision of the network?

Given a new sample $\mathbf{x}_*$, we want to find the gradients that correspond to the input features, denoted by $\boldsymbol{\delta}^0$. Taking a similar approach as in Section 2.2, the approximate predictive distribution for the gradients of the input features is given by

$$q(\boldsymbol{\delta}^0|\mathbf{x}_*) = \int p(\boldsymbol{\delta}^0|\mathbf{x}_*, \boldsymbol{\theta})q(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{3}$$

Calculating $p(\boldsymbol{\delta}^0|\mathbf{x}_*,\ \boldsymbol{\theta})$ is done through the backpropagation algorithm, i.e. computing the gradients with respect to the output of the network and then using the chain rule to work backward toward the input gradients. Also, we modify the backward pass such that negative gradients are canceled, following the Guided Backpropagation procedure. For clear notation, we denote this procedure as $\nabla_{\boldsymbol{\theta}} f^{gb}(\mathbf{x}_*;\ \boldsymbol{\theta})$, where $\nabla_{\boldsymbol{\theta}}$ indicate finding the gradients of each layer with respect to the parameters of the network and $f^{gb}(\mathbf{x}_*;\ \boldsymbol{\theta})$ is the prediction of the model with the modified backward pass. The predictive distribution in Eq. (1) can then be approximated using Monte Carlo integration as follows:

$$q(\boldsymbol{\delta}^0|\mathbf{x}_*) = \frac{1}{T}\sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}} f^{gb}(\mathbf{x}_*; \mathbf{W}_t^*). \tag{4}$$

In practice, this amounts to performing T forward and backward passes with Dropout applied and computing the standard deviation over the gradients of each input pixel over all T samples. We refer to this method of estimating gradient uncertainty as Monte Carlo Guided Backpropagation.

3. Experiments

3.1. Experimental setup

We evaluate our methods on a recent benchmark dataset for polyp segmentation, namely the EndoScene dataset (Vázquez et al., 2016), which consists of 912 RGB images obtained from colonoscopies of 36 patients. Each input image has a corresponding annotated (labeled) image provided by physicians, where pixels belonging to a polyp are marked in white and pixels belonging to the colon are marked in black. We consider the binary task of classifying each pixel as polyp or part of the colon (background class). Following the approach of Vázquez et al. (2016) we separate the dataset into a training, validation, and test set. The training set consists of 20 patients and 547 images, the validation set consists of 8 patients and 183 images, and the test set consists of 8 patients and 182 images. All RGB input images are normalized to the range [0,1]. All models were trained using ADAM (Kingma and Ba, 2014)

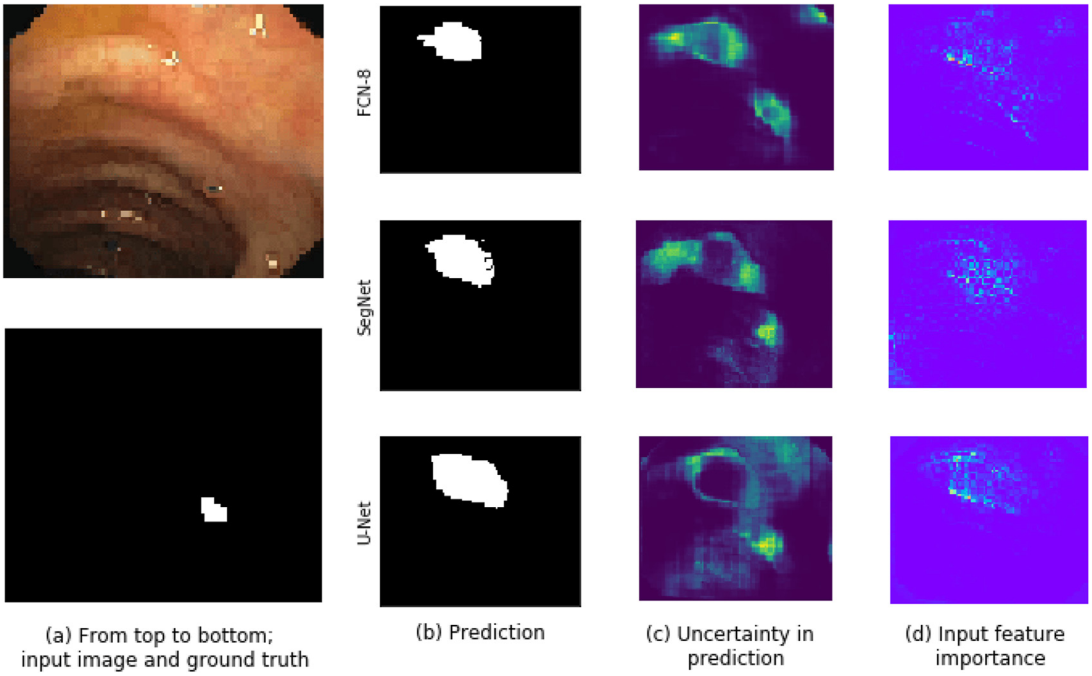

FCN-8
SegNet
U-Net
(a) From top to bottom; input image and ground truth
(b) Prediction
(c) Uncertainty in prediction
(d) Input feature importance

Fig. 4. Figure displays the prediction, uncertainty map, and interpretability map for the FCN-8, SegNet and U-Net, for the input image shown in the leftmost column. Best viewed in color.

Table 1
Results on the EndoScene test dataset.

Model	# Parameters(M)	IoU background	IoU polyp	Mean IoU	Global Accuracy
SDEM (Bernal et al., 2014)	-	0.799	0.221	0.412	0.756
U-Net	27.5	0.945	0.516	0.723	0.945
SegNet	29.5	0.933	0.522	0.727	0.935
FCN-8 (Vázquez et al., 2016)	134.5	**0.946**	0.509	0.727	**0.949**
FCN-8	134.5	**0.946**	**0.587**	**0.767**	**0.949**

with a batch size of 10 and a cross-entropy loss. We use the validation set to apply early stopping by monitoring the polyp IoU score with a patience of 30. For performance evaluation, we calculate the Intersection over Union (IoU) metric and global accuracy (per-pixel accuracy) on the test set. For a given class c, prediction $\hat{y}_i$ and ground truth y_i, the IoU is defined as

$$\mathrm{IoU}(c) = \frac{\sum_i (\hat{y}_i == c \wedge y_i == c)}{\sum_i (\hat{y}_i == c \vee y_i == c)} \tag{5}$$

where $\wedge$ is the logical *and* operation and $\vee$ is the logical *or* operation.

Additionally, we evaluated our proposed method for estimating uncertainty in input feature importance on the 2015 MICCAI polyp detection challenge (Bernal et al., 2017). As the test images of this dataset are of high quality and our proposed approach is mostly a visual technique, assessing our method on this data will provide further validation of our method.

3.2. Quantitative and qualitative results

*Quantitative results*In Table 1 we report our results for the FCN-8, SegNet and U-Net along with the results of previous works on polyp segmentation from both traditional machine learning and deep learning based approaches. The traditional machine learning method computes a histogram based on the pixel values and uses peaks and valleys information from the histogram to perform segmentation. It is referred to as the Segmentation from Energy Maps (SDEM) algorithm (Bernal et al., 2014). For the deep learning approach, segmentation is performed using the FCN-8, but without Batch Normalization or transfer learning. This approach is referred to as FCN-8 in Table 1. The results show that all deep learning approaches significantly outperform the more traditional machine learning approach, and the difference in performance between our implementation of the FCN-8 and that of Vázquez et al. (2016) demonstrates that including recent advances in deep learning methodology can improve performance.

Qualitative results Fig. 2(b) and 4(b) displays some qualitative results on the test data for the FCN-8, SegNet and U-Net. Fig. 2 shows a typical example where a large, elliptical polyp is located with high precision by all three models. In Fig. 4 we present a more challenging example where all models fail to locate the small polyp present in the image. Interested readers can find additional results in Appendixs B and C.

3.3. Modeling uncertainty in prediction

Figs. 2(c) and 4(c) present examples of uncertainty estimation for the FCN-8, SegNet and U-Net, respectively, using Monte Carlo Dropout. These uncertainty maps are obtained by sampling 10 predictions from each model with a dropout rate of 0.5 and estimating the standard deviation for each pixel. Pixels displayed in bright green are associated with high uncertainty while pixels displayed in dark blue are associated with low uncertainty.

The example shown in Fig. 2 shows that all models have high confidence for most pixels in their prediction, with the exception of pixels around the border of the polyp itself. This is reasonable, as it is difficult to assess exactly where the polyp starts and the colon ends. In the example shown in Fig. 4, where all models make

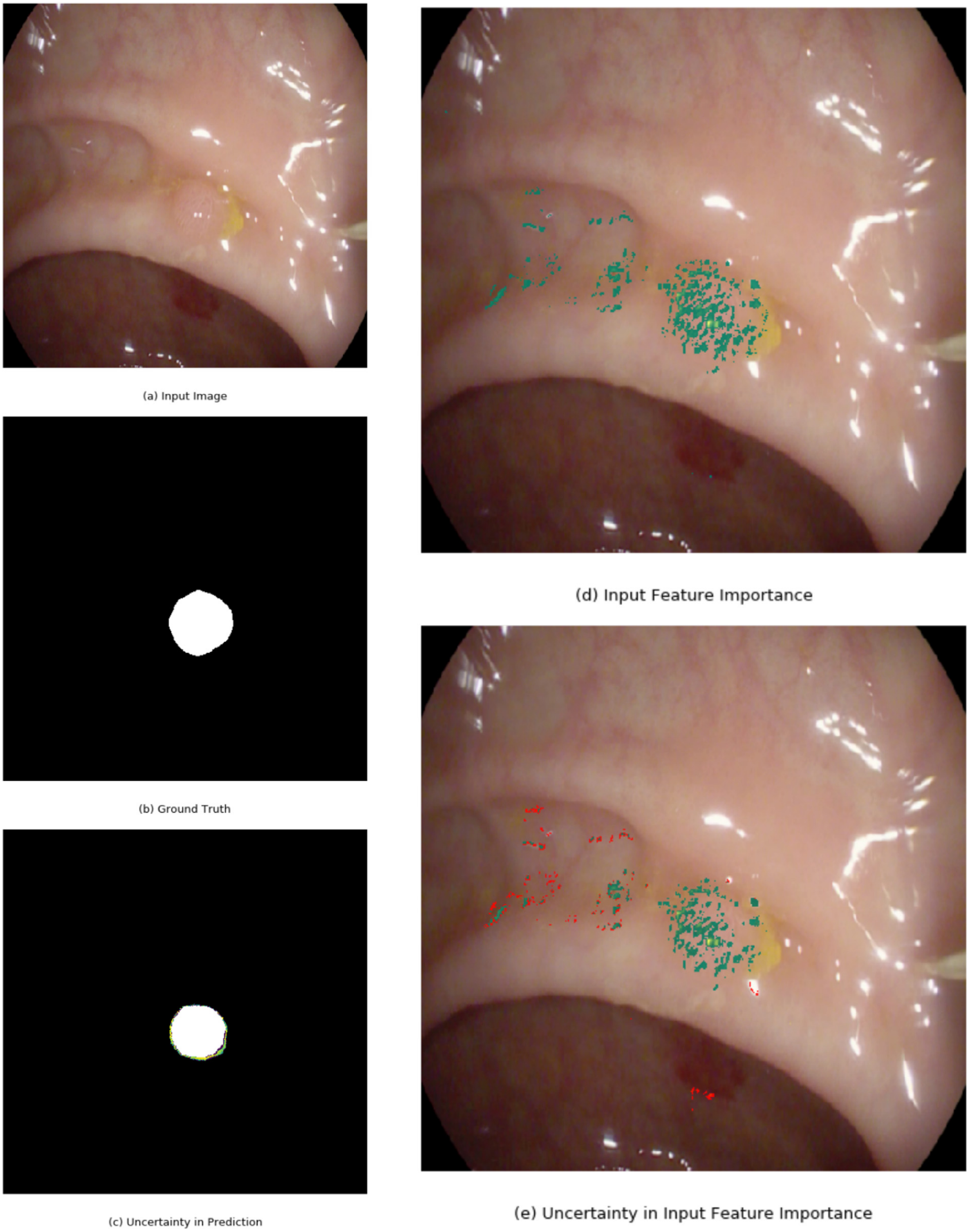

(a) Input Image
(b) Ground Truth
(c) Uncertainty in Prediction
(d) Input Feature Importance
(e) Uncertainty in Input Feature Importance

Fig. 5. Figure displays input image (a), ground truth (b), prediction with uncertainty overlaid (c), input feature importance (d), and uncertainty in input feature importance (e). For the uncertainty in input feature importance results, pixels colored green indicate that the features are important for the prediction of polyps and that the model is certain of its importance. Pixels colored red indicate features that might be important for the prediction of polyps but the model is uncertain of its importance. Best viewed in color. Input image originated from the MICCAI dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

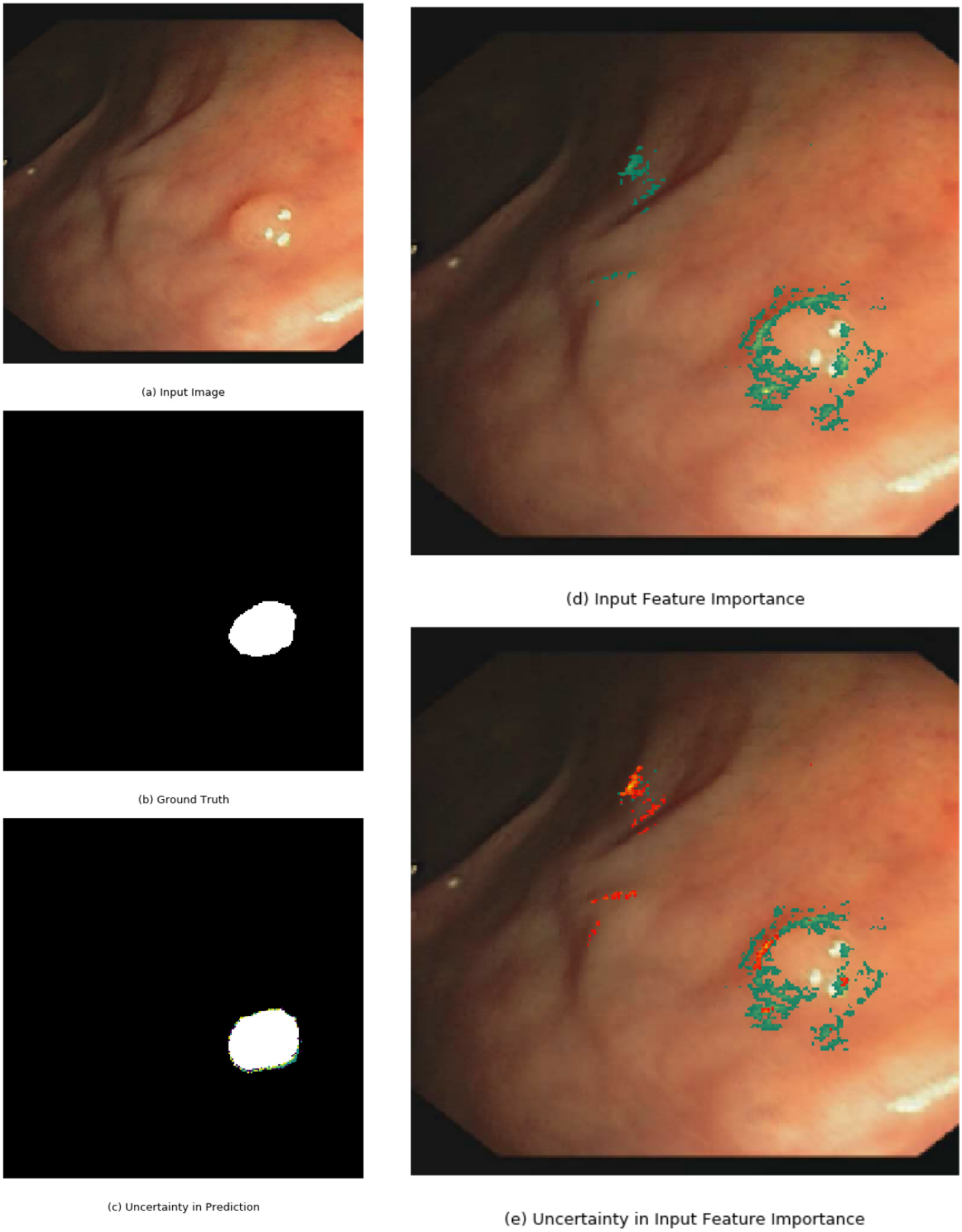
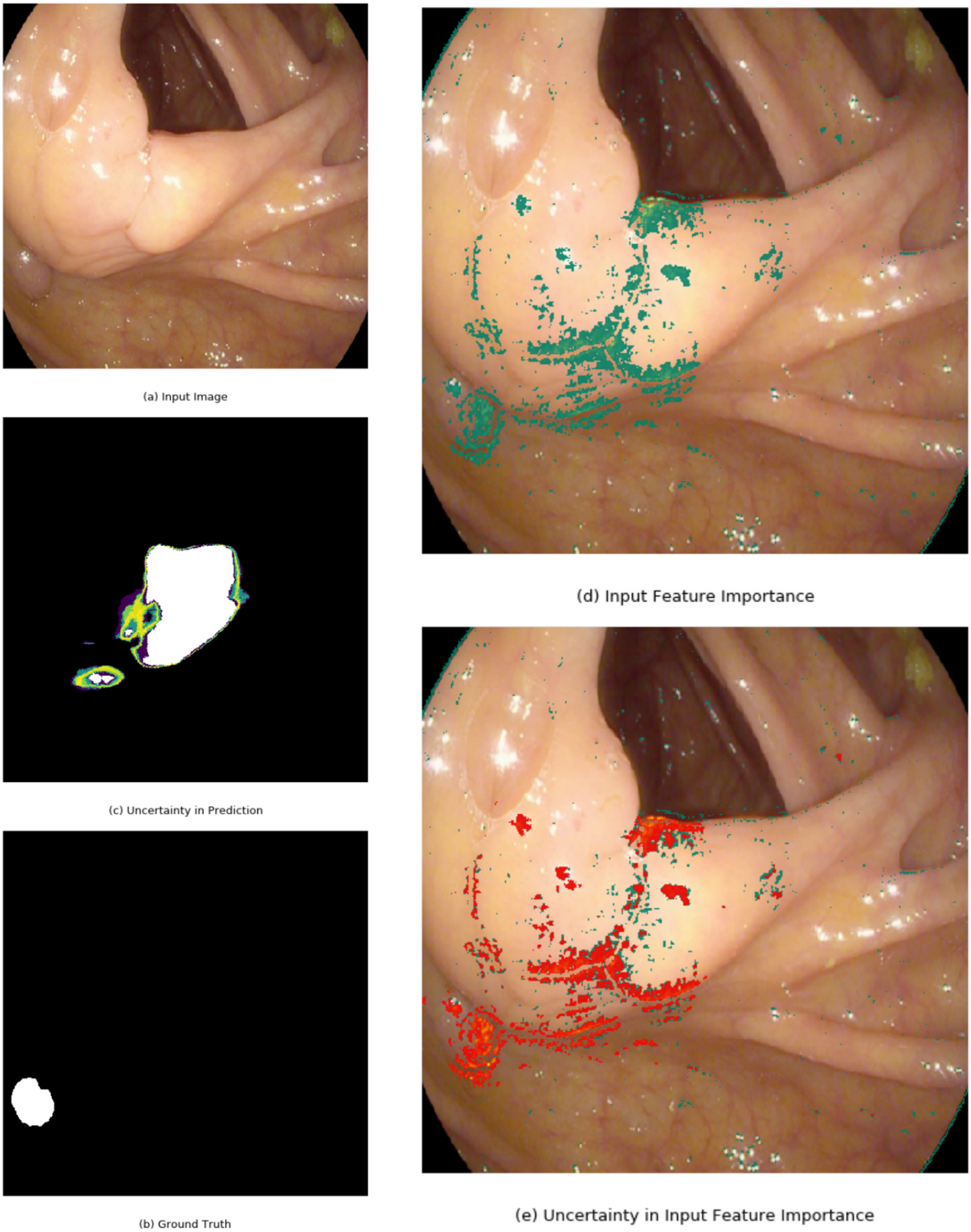

(a) Input Image
(b) Ground Truth
(c) Uncertainty in Prediction
(d) Input Feature Importance
(e) Uncertainty in Input Feature Importance

Fig. 6. Figure displays input image (a), ground truth (b), prediction with uncertainty overlaid (c), input feature importance (d), and uncertainty in input feature importance (e). For the uncertainty in input feature importance results, pixels colored green indicate that the features are important for the prediction of polyps and that the model is certain of its importance. Pixels colored red indicate features that might be important for the prediction of polyps but the model is uncertain of its importance. Best viewed in color. Input image originated from the Endoscene dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)


(a) Input Image

(c) Uncertainty in Prediction

(b) Ground Truth

(d) Input Feature Importance

(e) Uncertainty in Input Feature Importance

Fig. 7. Figure displays input image (a), ground truth (b), prediction with uncertainty overlaid (c), input feature importance (d), and uncertainty in input feature importance (e). For the uncertainty in input feature importance results, pixels colored green indicate that the features are important for the prediction of polyps and that the model is certain of its importance. Pixels colored red indicate features that might be important for the prediction of polyps but the model is uncertain of its importance. Best viewed in color. Input image originated from the MICCAI dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

inaccurate predictions, the uncertainty estimates look notably different, with large regions of uncertainty for all three models. The examples shown in Figs. 2 and 4 demonstrate how seemingly similar predictions can have different uncertainty estimates for the different types of networks investigated in this work, and that erroneous predictions show distinctively different uncertainty estimates than correct predictions.

Fig. 3 displays how precision and recall is related to uncertainty in predictions. It shows the overall precision and recall for each class on the Endoscene test dataset when pixel with a mean-class uncertainty above a certain threshold are excluded. The estimated uncertainty for each class have been normalized into values between 0 and 1. Results in Fig. 3 (a) display how precision decreases as more pixel predictions with high uncertainty are included. This connection between precision and uncertainty agrees with the qualitative examples in Figs. 2 and 4 discussed above. Results in Fig. 3 (b) show how recall slightly increases for the polyp class at a low uncertainty threshold, but then remains unchanged for both classes. The interested reader can find a similar experiment on the MICCAI dataset in Appendix C.

3.4. Modeling input feature importance

Figs. 2 (d) and 4 (d) show examples where Guided Backpropagation has been used to analyze the FCN-8, SegNet and U-Net, respectively. Pixels displayed in bright green are associated with pixels that are important to the prediction of the model while pixels displayed in blue are associated with pixels that are less important to the final prediction.

Fig. 2 indicates that all models are considering the edges of the polyp to make their prediction, where particularly the leftmost and bottom edge of the polyp is highlighted as important by all models. Fig. 4, where all models fail to locate the polyp, displays more disagreement between the models as to what pixels are important.

3.5. Modeling uncertainty in input feature importance

In order to focus on the new methodology we only use one model to evaluate our proposed method. The overall best performing segmentation model, FCN-8, was chosen to evaluate the proposed methodology for estimating uncertainty in input feature importance and demonstrate its merit. Figs. 5–7 presents examples of uncertainty estimation for input feature importance for the FCN-8 using Monte Carlo Guided Backpropagation. These results are obtained by sampling 10 gradient estimates from each model with a dropout rate of 0.5. The figures display: (a) the input image; (b) the ground truth; (c) prediction with uncertainty overlaid; (d) input feature importance; and (e) uncertainty in input feature importance. For the uncertainty in input feature importance results, pixels colored green indicate that the features are important for the prediction of polyps and that the model is certain of its importance. Pixels colored red indicate features that might be important for the prediction of polyps but the model is uncertain of its importance. Examples shown in Figs. 5 and 7 are from the test set of the MICCAI dataset while the example shown in Fig. 6 is from the test set of the Endoscene dataset. Interested readers can find additional examples of uncertainty estimation for input feature importance in Appendix B.

Fig. 5 displays an example where the FCN-8 makes a successful segmentation. The interpretability map in Fig. 5 (d) indicates that there are two regions of importance in the input image, one corresponding to the polyp and one region towards the leftmost part of the image. However, the uncertainty in the input feature importance map in Fig. 5 (e) shows that the model is uncertain of the leftmost feature's importance, while the features corresponding to the polyp itself have a high degree of certainty.

Fig. 6 shows another example where the FCN-8 makes a successful segmentation, but also highlight important input features towards the leftmost part of the image, in addition to the polyp itself. Fig. 6 (e) displays that the FCN-8 is highly confident in the importance of the features corresponding to the polyp itself, but indicate a high degree of uncertainty for the highlighted regions towards the leftmost part of the image.

Fig. 7 exhibits an example from the MICCAI dataset where the FCN-8 fails to locate the polyp present in the image, but instead segments a large portion of the colon as polyp. While the interpretability maps in Fig. 7 (d) show large regions of important pixels, it is evident from Fig. 7 (e) that none of the regions have a high degree of importance. As the prediction with uncertainty overlayed in Fig. 7 (e) also indicates regions of uncertainty, practitioners would be wary to trust the model's prediction in this case.

4. Conclusion

In this work we have demonstrated how DSSs based on deep learning can be interpretable and provide uncertainty estimates with their predictions. Moreover, we presented a novel method for estimating uncertainty in input feature importance and demonstrated how this technique can be used to model uncertainty in input pixel importance. Our results demonstrate that the models considered in these experiments exploit edge and shape information of polyps in order to make their predictions and that uncertainty differs significantly between false and correct predictions.

Declaration of Competing Interest

All authors declare that they have no conflicts of interest regarding the publication of this paper.

Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPU used for this research.

Appendix A. Network details

In order to perform per pixel predictions, FCNs employ an encoder-decoder architecture and are capable of end-to-end learning. The encoder network extracts useful features from an image and maps it to a low-resolution representation. The decoder network is tasked with mapping the low-resolution representation back into the same resolution as the input image. Upsampling in FCNs is performed using a fixed upsampling approach, like bilinear or nearest neighbor interpolation, or by learning the upsampling procedure as part of the model optimization via transposed convolutions. Learned upsampling filters add additional parameters to the network architecture, but tend to provide better overall results (Shelhamer et al., 2017). Upsampling can further be improved by including skip connections, which combine coarse level semantic information with higher resolution segmentation from previous network layers. Due to the lack of fully connected layers, inference can be performed on images of arbitrary size.

A1. FCN-8

The FCN-8 was introduced by Shelhamer et al. (2017) and consists of an encoder network and a decoder network, where the encoder network is based on the VGG-16 architecture (Simonyan and Zisserman, 2015) and consists of five encoders. The decoder network consists of three decoders. Dropout (Srivastava et al., 2014), a regularization technique that randomly set units in a layer to zero, is included between all layers of the first decoder. Upsampling is

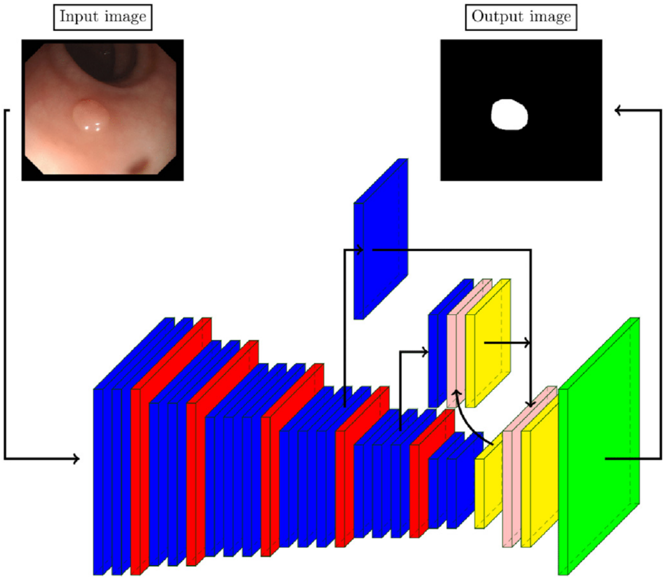


Fig. A.8. An illustration of the FCN-8. Color codes description: Blue - Convolution (3x3), Batch Normalization and ReLU; Yellow - Upsampling; Pink - Summing; Red - Pooling (2x2); Green - Soft-max. Dropout was included as proposed by Simonyan and Zisserman (2015) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.).

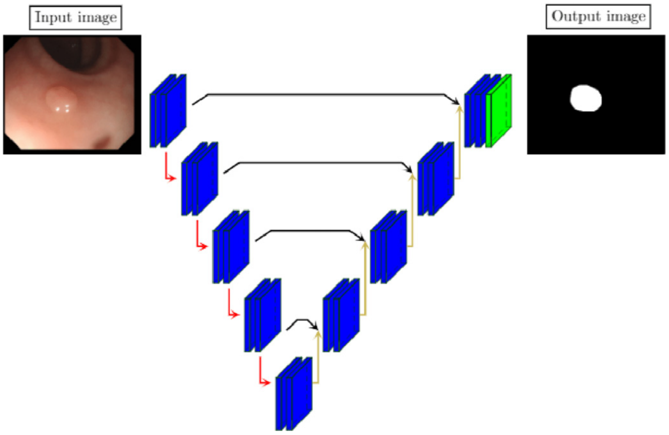


Fig. A.9. An illustration of the U-Net. Color codes description: Blue - Convolution (3x3), Batch Normalization and ReLU; Green - Soft-max; Yellow arrow - Upsampling; Black arrow - Concatenate; Red arrow - Pooling (2x2) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.).

performed using transposed convolutions at the end of each encoder and skip connections are included between the three central encoders and the decoders. Note that we have added Batch Normalization (Ioffe and Szegedy, 2015) in our implementation and that the encoder weights are initialized with pretrained weights from a VGG16 model (Simonyan and Zisserman, 2015) that was previously trained on the ImageNet dataset (Deng et al., 2009).

A2. U-Net

One of the first networks to build upon FCNs was the U-Net (Ronneberger et al., 2015), which is comprised of an encoder network consisting of five encoders and a decoder network consisting of four decoders. U-Net introduced an alternative method to recover the resolution of the data where the feature maps produced in the fifth encoder is upsampled by a factor of two using transposed convolution and concatenated with the feature maps produced by the fourth encoder. These combined feature maps are passed into the first decoder, which in turn is upsampled and concatenated with the feature maps of the third encoder. This process is repeated until the resolution of the input feature map is recovered. The final decoder is followed by a 1 × 1 convolutions that maps the feature vector into the desired number of classes and a softmax function. Dropout is applied after each layer of the final encoder. We included Batch Normalization after each layer, except for layers preceding a transposed convolution and the final layer.

A3. SegNet

Both the FCN-8 and the U-Net rely on transposed convolutions to recover feature maps with the same resolution as the input features. SegNet (Badrinarayanan et al., 2017), instead, presents another option and is made up of a symmetrically structured encoder decoder network, where the encoder network consists of five encoders based on the VGG-16 (Simonyan and Zisserman, 2015) and the decoder consists of five decoders. The decoder network is identical to the encoder network but with the max-pooling operation replaced by a max-unpooling operation. When a feature map is downsampled the max-pooling indices are stored and used at a later stage to perform non-linear upsampling, a procedure with several advantages. Firstly, it produces sparse feature maps that are computationally attractive and implicit feature selectors. Secondly, it removes the need to learn additional filter for upsampling, thus reducing the number of parameters in the model. Dropout was included after the three central encoders and decoders inspired by Kendall et al. (2015).

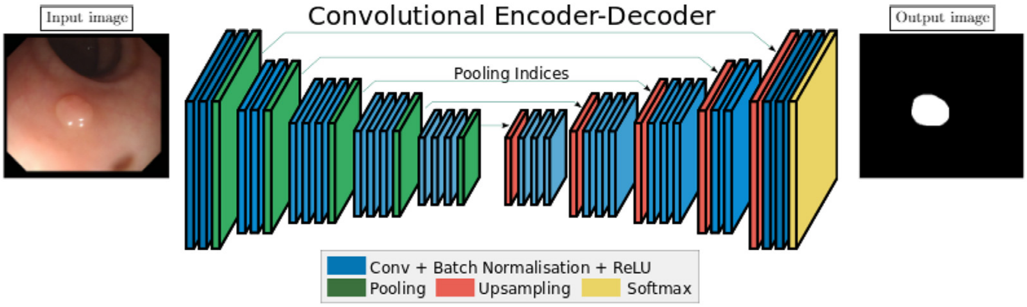

Input image
Convolutional Encoder-Decoder
Pooling Indices
Output image
Conv + Batch Normalisation + ReLU
Pooling
Upsampling
Softmax

Fig. A.10. An illustration of SegNet, originally obtained from Badrinarayanan et al. (2017). Color codes description: Blue - Convolution (3x3), Batch Normalization and ReLU; Green - Soft-max; Yellow arrow - Upsampling; Black arrow - Concatenate; Red arrow - Pooling (2x2) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.).

Appendix B. additional qualitative results

Figs. B.11–B.13 display additional results on test images from the Endoscene dataset for the FCN-8, SegNet and U-Net, respectively. Each row represents, from top to bottom, input image, ground truth, prediction, uncertainty map, and interpretability map. Results were obtained using the same procedure as described in the main paper.

Figs. B.14–B.16 display additional results of estimating uncertainty in input feature importance for the FCN-8. These results are also obtained following the same procedure described in the main paper.

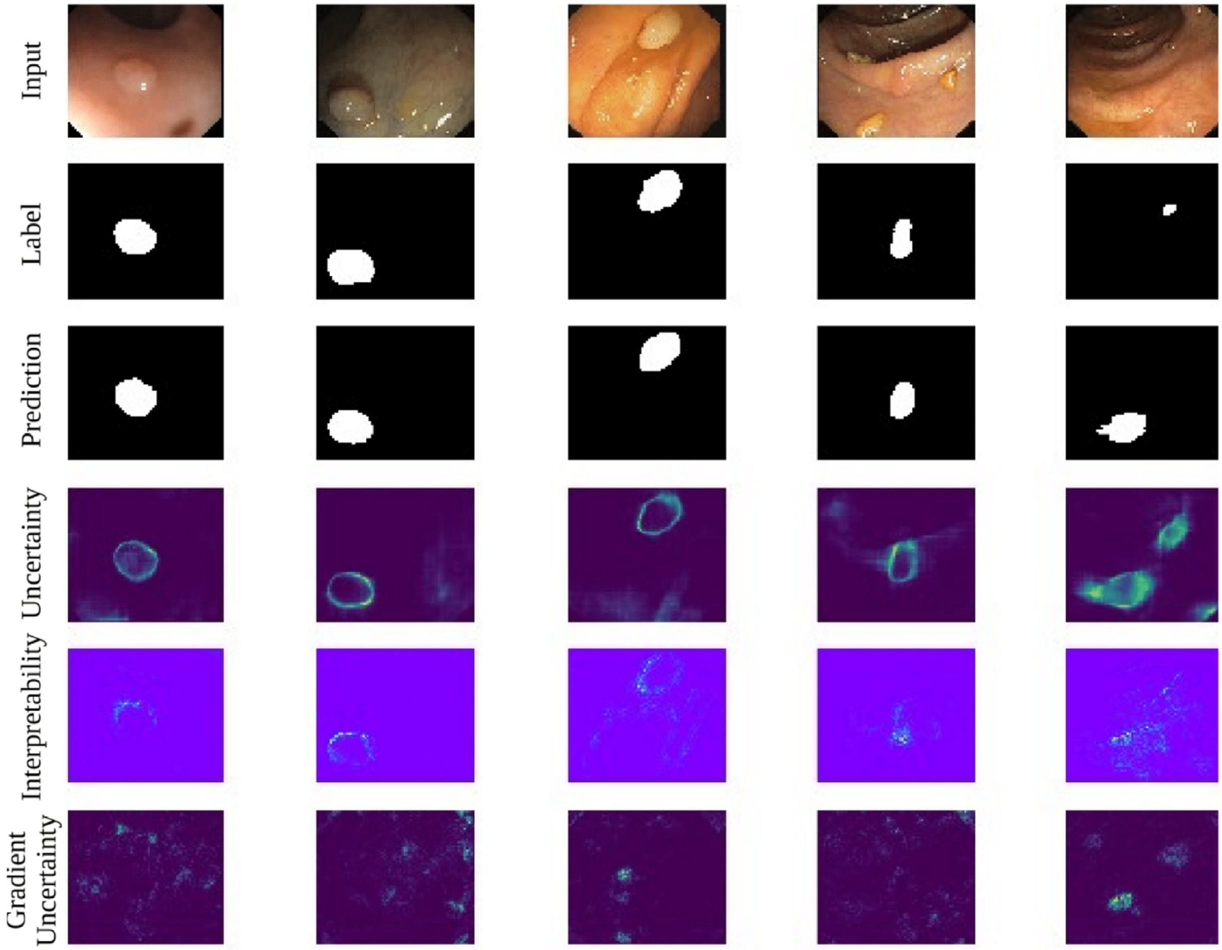


Fig. B.11. Figure displays FCN-8's predictions, the uncertainty map associated with the predictions, and the input features the network deems important. Each row represents, from top to bottom, input image, ground truth, prediction, uncertainty map, and interpretability map. White pixels are classified as polyps and black pixels are classified as background class. For the uncertainty maps, dark blue pixels are associated with low uncertainty and bright green pixels are associated with high uncertainty. For the interpretability maps, bright green pixels are considered important to the prediction of the network. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

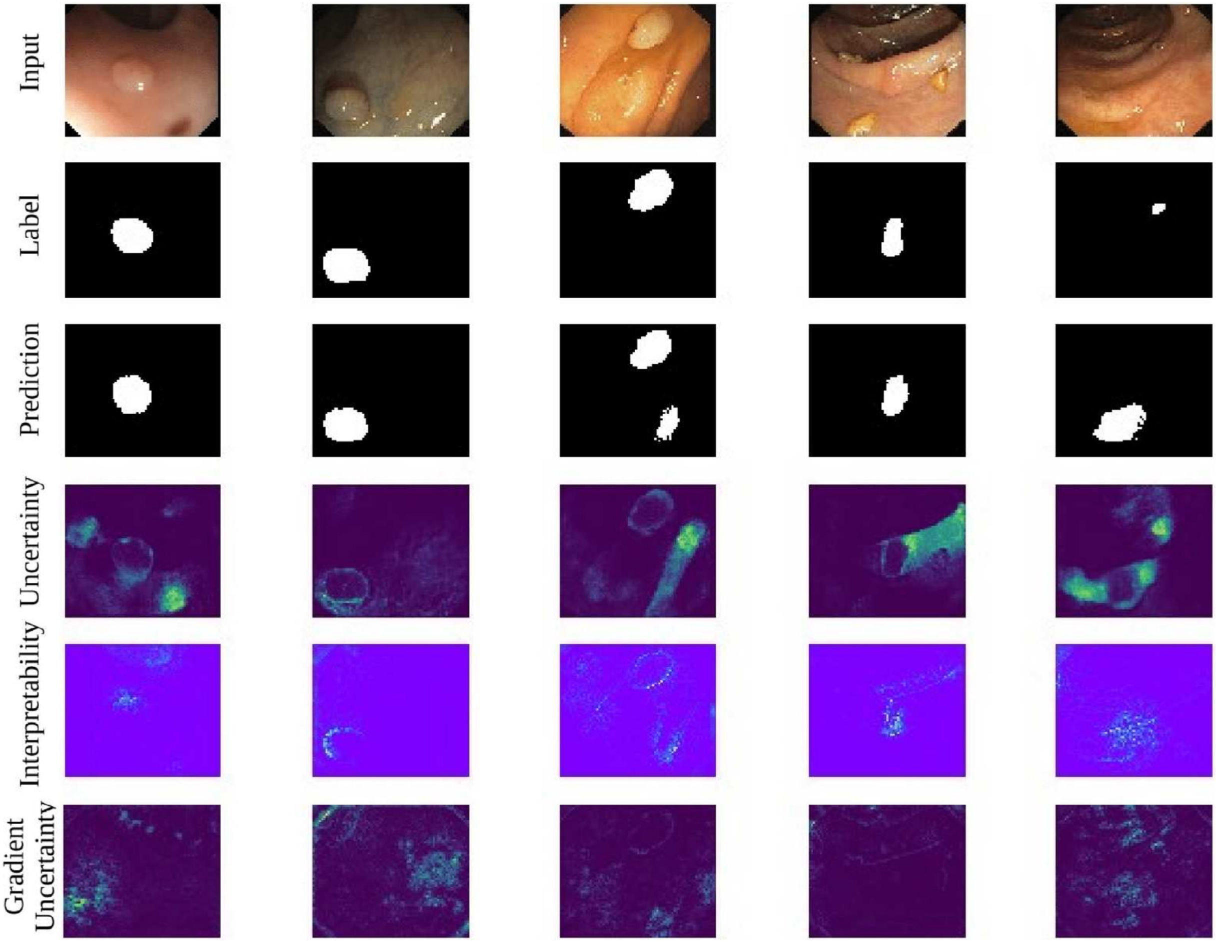


Fig. B.12. Figure displays SegNet's predictions, the uncertainty map associated with the predictions, and the input features the network deems important. Each row represents, from top to bottom, input image, ground truth, prediction, uncertainty map, and interpretability map. White pixels are classified as polyps and black pixels are classified as background class. For the uncertainty maps, dark blue pixels are associated with low uncertainty and bright green pixels are associated with high uncertainty. For the interpretability maps, bright green pixels are considered important to the prediction of the network. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

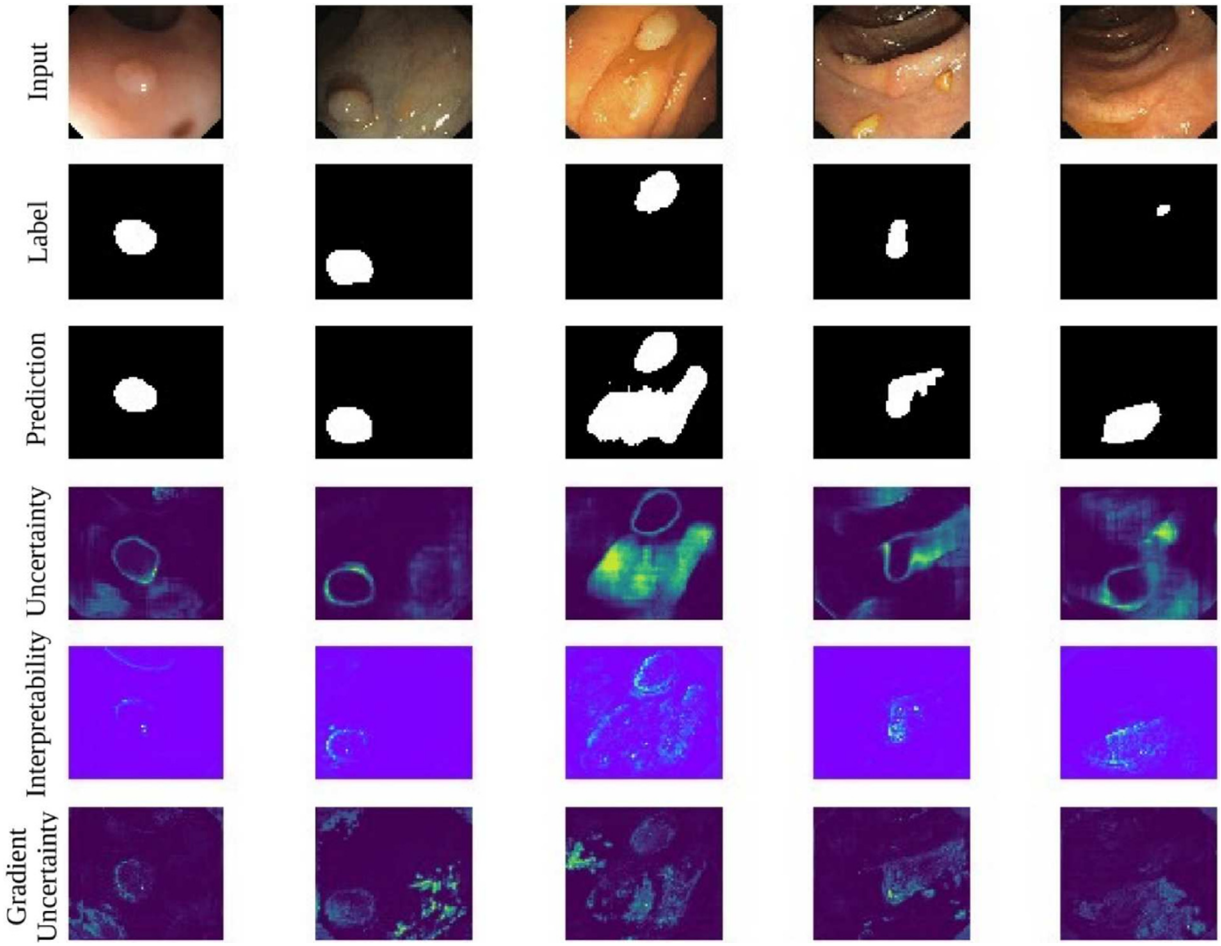
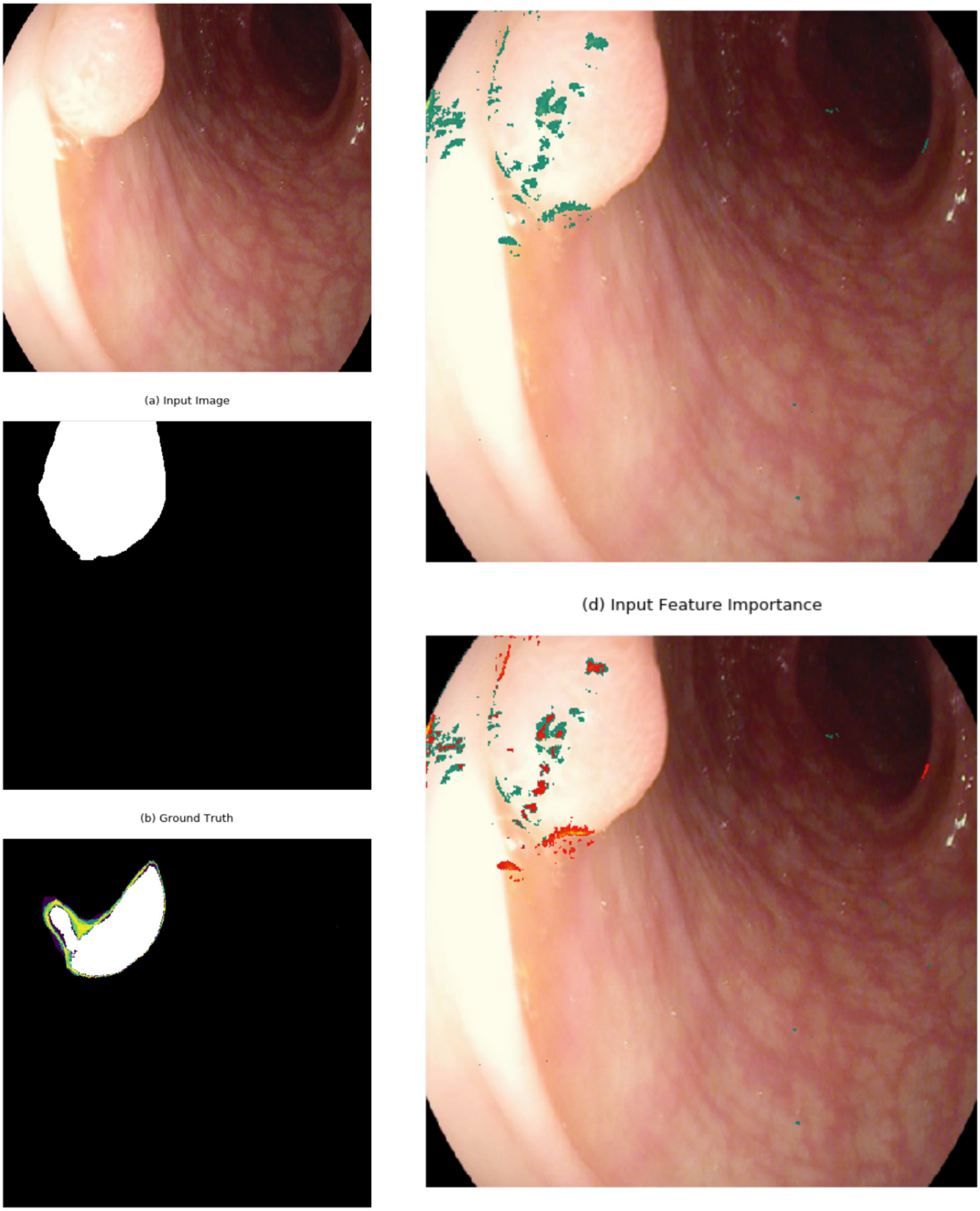


Fig. B.13. Figure displays U-Net's predictions, the uncertainty map associated with the predictions, and the input features the network deems important. Each row represents, from top to bottom, input image, ground truth, prediction, uncertainty map, and interpretability map. White pixels are classified as polyps and black pixels are classified as background class. For the uncertainty maps, dark blue pixels are associated with low uncertainty and bright green pixels are associated with high uncertainty. For the interpretability maps, bright green pixels are considered important to the prediction of the network. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



(a) Input Image

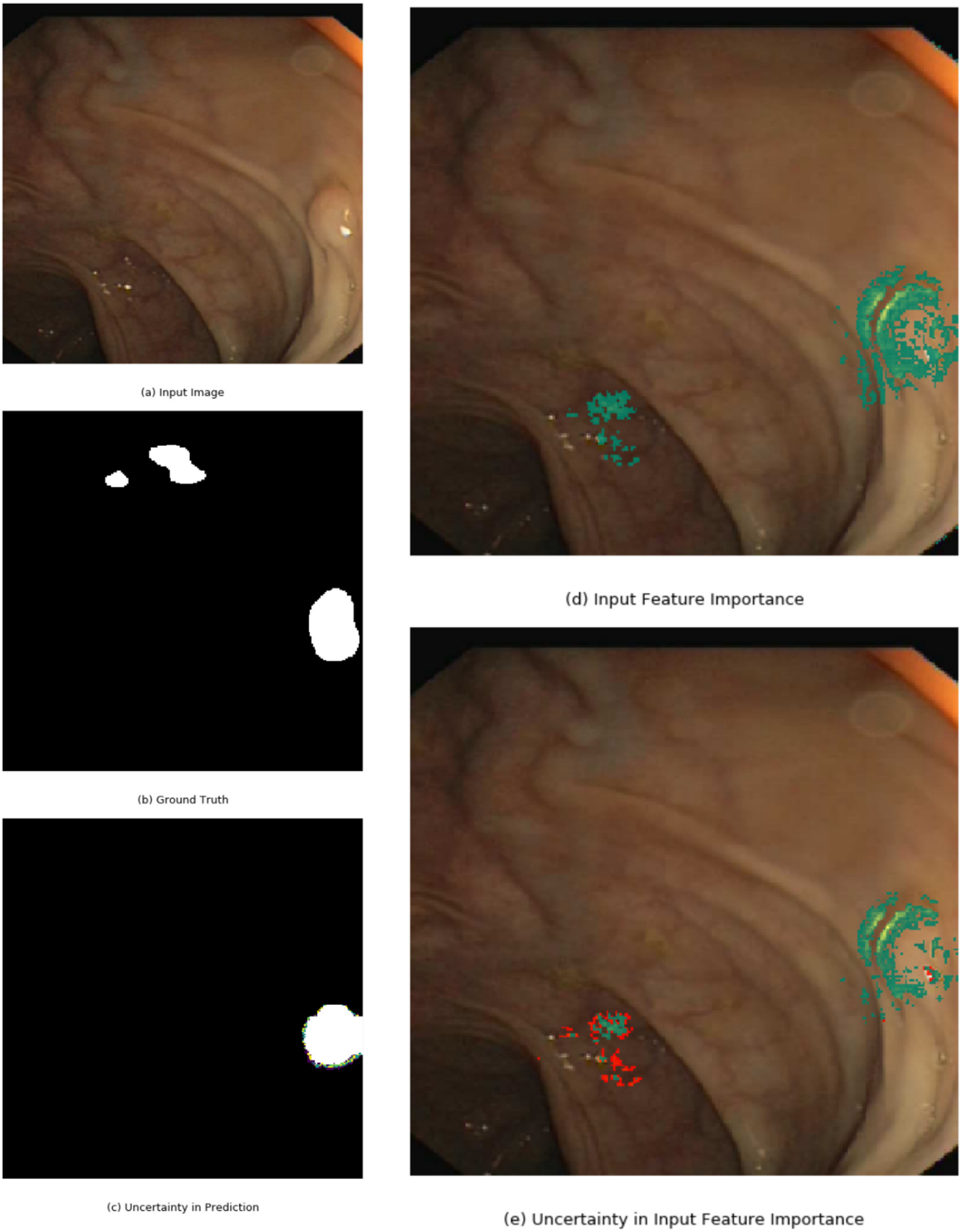
(b) Ground Truth

(c) Uncertainty in Prediction

(d) Input Feature Importance

(e) Uncertainty in Input Feature Importance

Fig. B.14. Figure displays input image (a), ground truth (b), prediction with uncertainty overlaid (c), input feature importance (d), and uncertainty in input feature importance (e). For the uncertainty in input feature importance results, pixels colored green indicate that the features are important for the prediction of polyps and that the model is certain of its importance. Pixels colored red indicate features that might be important for the prediction of polyps but the model is uncertain of its importance. Best viewed in color. Input image originated from the MICCAI dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



(a) Input Image

(b) Ground Truth

(c) Uncertainty in Prediction

(d) Input Feature Importance

(e) Uncertainty in Input Feature Importance

Fig. B.15. Figure displays input image (a), ground truth (b), prediction with uncertainty overlaid (c), input feature importance (d), and uncertainty in input feature importance (e). For the uncertainty in input feature importance results, pixels colored green indicate that the features are important for the prediction of polyps and that the model is certain of its importance. Pixels colored red indicate features that might be important for the prediction of polyps but the model is uncertain of its importance. Best viewed in color. Input image originated from the Endoscene dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

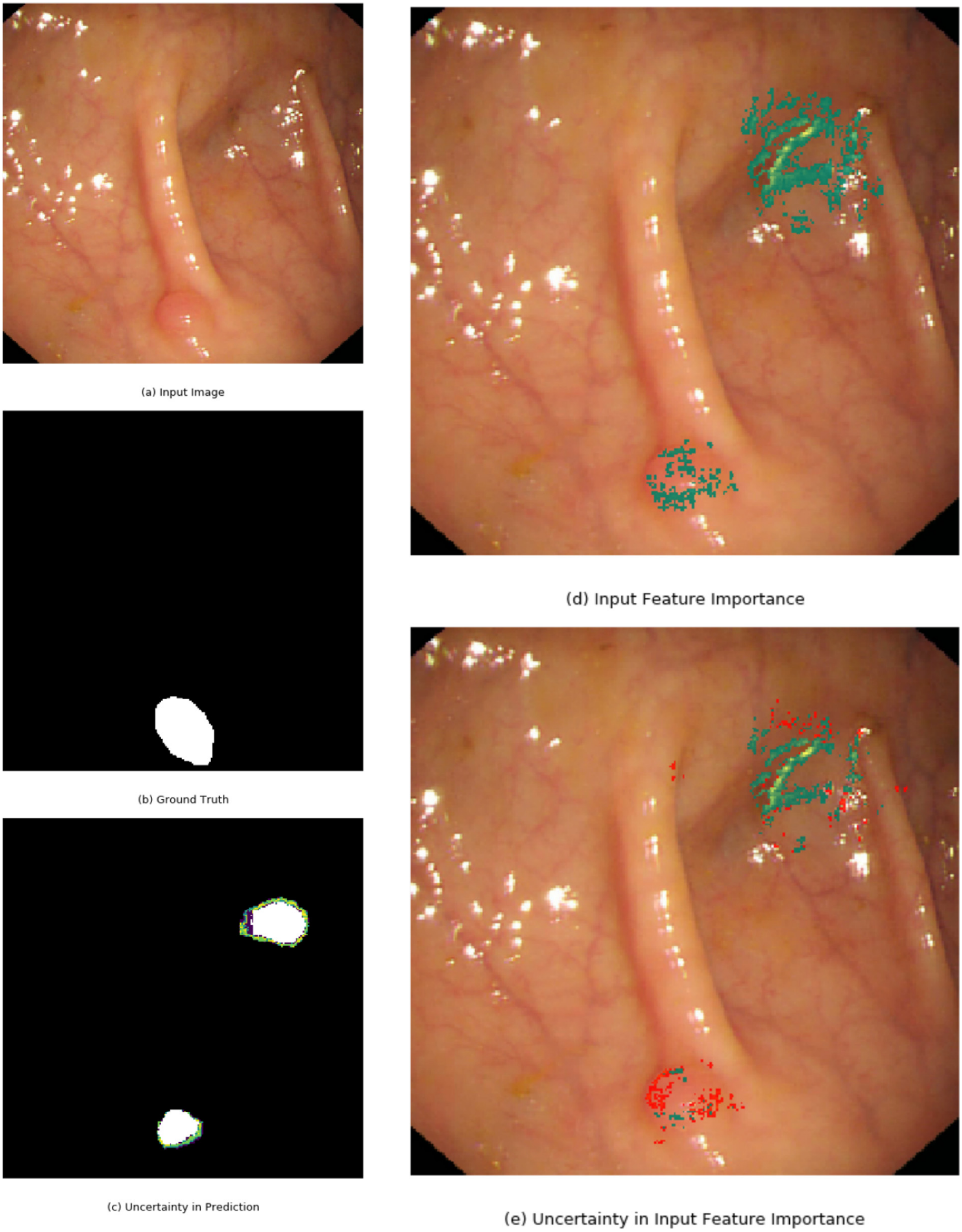

(a) Input Image
(b) Ground Truth
(c) Uncertainty in Prediction
(d) Input Feature Importance
(e) Uncertainty in Input Feature Importance

Fig. B.16. Figure displays input image (a), ground truth (b), prediction with uncertainty overlaid (c), input feature importance (d), and uncertainty in input feature importance (e). For the uncertainty in input feature importance results, pixels colored green indicate that the features are important for the prediction of polyps and that the model is certain of its importance. Pixels colored red indicate features that might be important for the prediction of polyps but the model is uncertain of its importance. Best viewed in color. Input image originated from the Endoscene dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Appendix C. Additional Qualitative Results on MICCAI dataset

Fig. C.17 and C.18 display additional results on test images from the MICCAI dataset for the FCN-8, SegNet and U-Net, respectively. Results were obtained using the same procedure as described in the main paper. Fig. C.19 displays how precision and recall is related to uncertainty in predictions on the MICCAI test data, similar to the experiment described in Section 3.3.

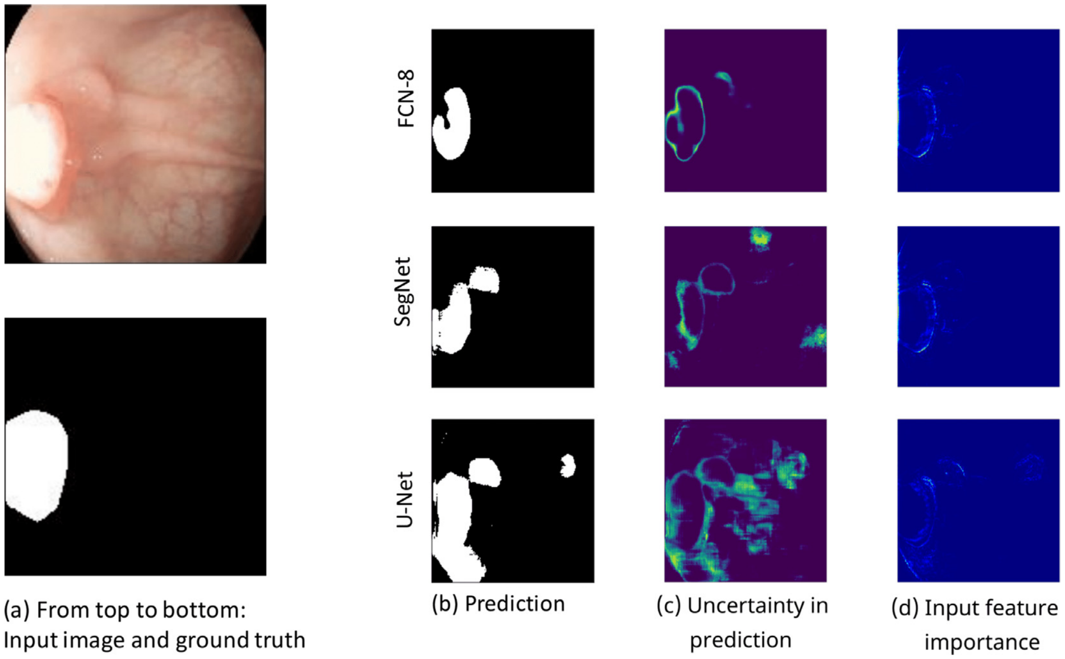

FCN-8
SegNet
U-Net
(a) From top to bottom: Input image and ground truth
(b) Prediction
(c) Uncertainty in prediction
(d) Input feature importance

Fig. C.17. Figure displays the prediction, uncertainty map, and interpretability map for the FCN-8, SegNet and U-Net, for the input image from the MICCAI dataset shown in the leftmost column. Best viewed in color.

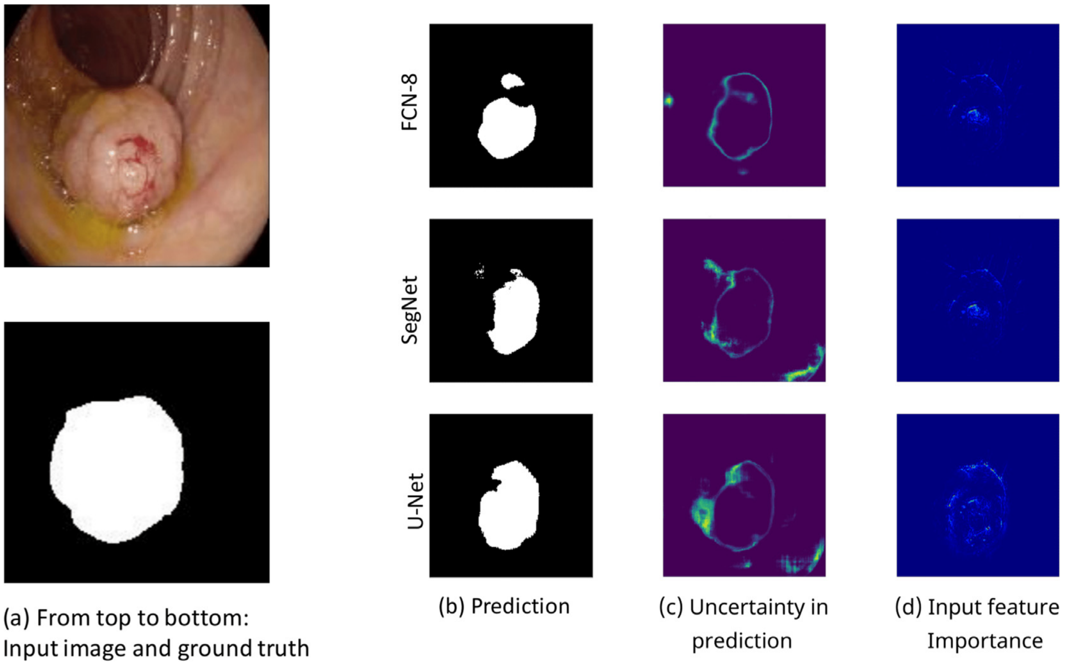
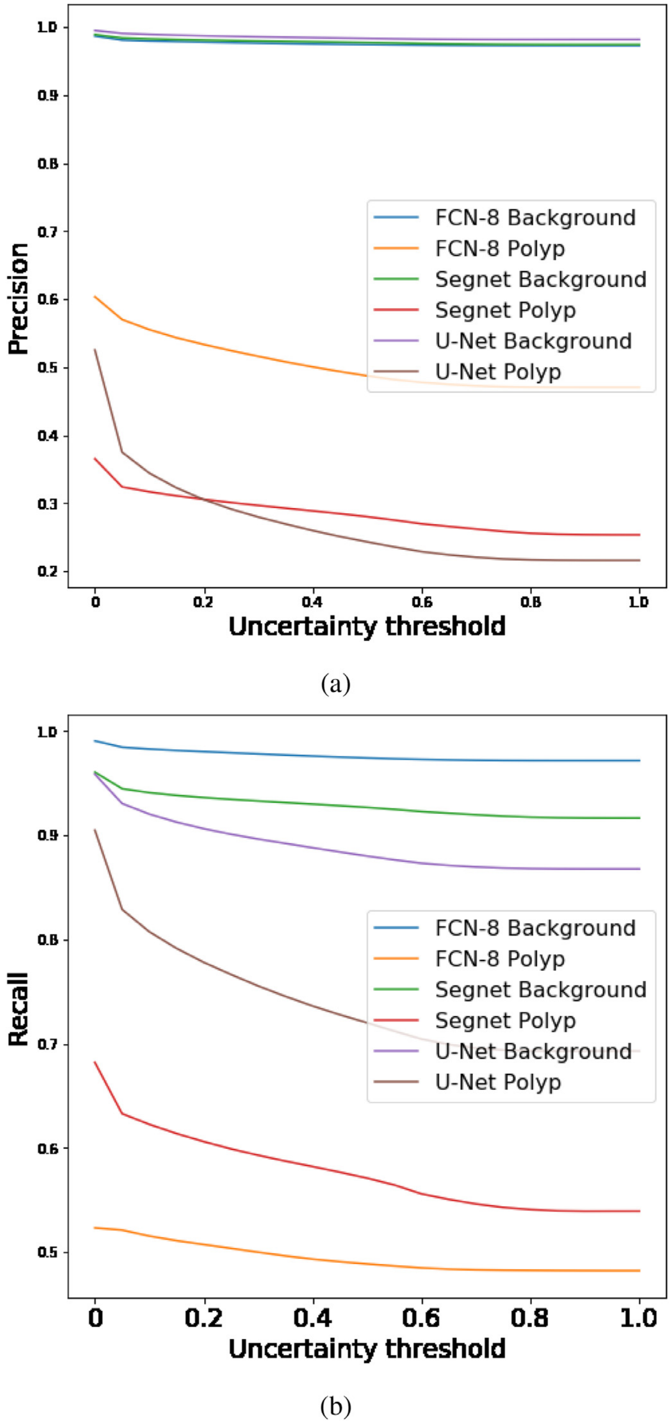

FCN-8
SegNet
U-Net
(a) From top to bottom: Input image and ground truth
(b) Prediction
(c) Uncertainty in prediction
(d) Input feature Importance

Fig. C.18. Figure displays the prediction, uncertainty map, and interpretability map for the FCN-8, SegNet and U-Net, for the input image from the MICCAI dataset shown in the leftmost column. Best viewed in color.



(a)

(b)

Fig. C.19. Precision and recall vs uncertainty plot for background and polyp class on the MICCAI test set.

References

Alain, G., Bengio, Y., 2017. Understanding intermediate layers using linear classifier probes. ArXiv: 1610.01644.

Alexandre, L.A., Casteleiro, J., Nobreinst, N., 2007. Polyp detection in endoscopic video using svms. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (Eds.), Knowledge Discovery in Databases: PKDD 2007. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 358–365.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., Samek, W., 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS One 10 (7), e0130140.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE TPAMI 2481–2495.

Bernal, J., Núñez, J.M., Sánchez, F.J., Vilariño, F., 2014. Polyp segmentation method in colonoscopy videos by means of Msa-Dova energy maps calculation. In: Workshop on Clinical Image-Based Procedures. Springer, pp. 41–49.

Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F., 2015. Wm-Dova maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. Comput. Med. Imaging Graph. 43, 99–111.

Bernal, J., Tajkbaksh, N., Sánchez, F.J., Matuszewski, B.J., Chen, H., Yu, L., Angermann, Q., Romain, O., Rustad, B., Balasingham, I., Pogorelov, K., Choi, S., Debard, Q., Maier-Hein, L., Speidel, S., Stoyanov, D., Brandao, P., Córdova, H., Sánchez-Montes, C., Gurudu, S.R., Fernández-Esparrach, G., Dray, X., Liang, J., Histace, A., 2017. Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge. IEEE Trans. Med. Imaging 36 (6), 1231–1249. doi:10.1109/TMI.2017.2664042.

Chen, W., Zheng, R., Baade, P.D., Zhang, S., Zeng, H., Bray, F., Jemal, A., Yu, X.Q., He, J., 2016. Cancer statistics in china, 2015. CA: A Cancer J. Clinic. 66 (2), 115–132. doi:10.3322/caac.21338.

Condessa, F., Bioucas-Dias, J., 2012. Segmentation and detection of colorectal polyps using local polynomial approximation. In: Campilho, A., Kamel, M. (Eds.), Image Analysis and Recognition. Springer Berlin Heidelberg, pp. 188–197.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A Large-Scale Hierarchical Image Database. In: Proceedings of the CVPR09, pp. 1097–1105.

Dubost, F., Adams, H., Bortsova, G., Ikram, M.A., Niessen, W., Vernooij, M., de Bruijne, M., 2019. 3D regression neural network for the quantification of enlarged perivascular spaces in brain mri. Med. Image Anal. 51, 89–100. doi:10.1016/j.media.2018.10.008.

Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: Proceedings of the ICML. JMLR.org, pp. 1050–1059.

Guo, S., Wang, K., Kang, H., Zhang, Y., Wang, K., Li, T., 2019. Bts-dsn: deeply supervised neural network with short connections for retinal vessel segmentation. Int. J. Med. Inf. doi:10.1016/j.ijmedinf.2019.03.015.

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

Hwang, S., Oh, J., Tavanapong, W., Wong, J., de Groen, P.C., 2007. Polyp detection in colonoscopy video using elliptical shape feature. In: Proceedings of the IEEE International Conference on Image Processing, 2. II–465–II–468. doi: 10.1109/ICIP.2007.4379193.

Häfner, M., Tamaki, T., Tanaka, S., Uhl, A., Wimmer, G., Yoshida, S., 2015. Local fractal dimension based approaches for colonic polyp classification. Med. Image Anal. 26 (1), 92–107. doi:10.1016/j.media.2015.08.007.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the ICML, pp. 448–456.

Kendall, A., Badrinarayanan, V., Cipolla, R., 2015. Bayesian segnet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. arXiv: 1511.02680.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision? In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.) Advances in Neural Information Processing Systems 30. Curran Associates, Inc., pp. 5574–5584.

Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization. arXiv: 1412.6980.

Larsen, I., 2016. Cancer in norway 2015 - cancer incidence, mortality, survival and prevalence in norway. oslo: Cancer registry of norway; 2016.

Liu, Q., 2017. Deep learning applied to automatic polyp detection in colonoscopy images : master thesis in system engineering with embedded systems.

Nida, N., Irtaza, A., Javed, A., Yousaf, M.H., Mahmood, M.T., 2019. Melanoma lesion detection and segmentation using deep region based convolutional neural network and fuzzy c-means clustering. Int. J. Med. Inf. 124, 37–48. doi:10.1016/j.ijmedinf.2019.01.005.

Brandao, P., Mazomenos, P., 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. Proc. SPIE 10134. 10134–10134–7. doi: 10.1117/12.2254361.

Ribeiro, E., Uhl, A., Häfner, M., 2016. Colonic polyp classification with convolutional neural networks. In: Proceedings of the IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS), pp. 253–258. doi:10.1109/CBMS.2016.39.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), MICCAI. Springer International Publishing, Cham, pp. 234–241.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1988. Neurocomputing: foundations of research. Nature 696–699.

Sharma, N., Ray, A., Shukla, K., Sharma, S., Pradhan, S., Srivastva, A., Aggarwal, L., 2010. Automated medical image segmentation techniques. J. Med. Phys. 35 (1), 3.

Shelhamer, E., Long, J., Darrell, T., 2017. Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (4), 640–651.

Shwartz-Ziv, R., Tishby, N., 2017. Opening the black box of deep neural networks via information. arXiv: 1703.00810.

Siegel, R.L., Miller, K.D., Jemal, A., 2017. Cancer statistics, 2017. CA: A Cancer J. Clinic. 67 (1), 7–30. doi:10.3322/caac.21387.

Simonyan, K., Vedaldi, A., Zisserman, A., 2013. Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv: 1312.6034.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. ICLR.

Springenberg, J., Dosovitskiy, A., Brox, T., Riedmiller, M., 2015. Striving for simplicity: The all convolutional net. In: Proceedings of the ICLR (Workshop track), p. N/A.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958.

Tajbakhsh, N., Gurudu, S.R., Liang, J., 2016. Automated polyp detection in colonoscopy videos using shape and context information. IEEE Trans. Med. Imaging 35 (2), 630–644. doi:10.1109/TMI.2015.2487997.

Urban, G., Tripathi, P., Alkayali, T., Mittal, M., Jalali, F., Karnes, W., Baldi, P., 2018. Deep learning localizes and identifies polyps in real time with 96 percent accuracy in screening colonoscopy. Gastroenterology doi:10.1053/j.gastro.2018.06.037.

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. Am. J. Gastroenterol. 101 (2), 343.

Vázquez, D., Bernal, J., Javier Sánchez, F., Fernández-Esparrach, G., López, A., Romero, A., Drozdzal, M., Courville, A., 2016. A benchmark for endoluminal scene segmentation of colonoscopy images. J. Healthcare Eng. 2017.

Werbos, P., 1974. Beyond regression: New tools for predicting and analysis in the behavioral sciences. Ph.D. thesis. Harvard University.

Wickstrøm, K., Kampffmeyer, M., Jenssen, R., 2018. Uncertainty modeling and interpretability in convolutional neural networks for polyp segmentation. In: Proceedings of the IEEE (MLSP), pp. 1–6. doi:10.1109/MLSP.2018.8516998.

Wimmer, G., Tamaki, T., Tischendorf, J., Häfner, M., Yoshida, S., Tanaka, S., Uhl, A., 2016. Directional wavelet based features for colonic polyp classification. Med. Image Anal. 31, 16–36. doi:10.1016/j.media.2016.02.001.

Yu, S., Príncipe, J.C., 2018. Understanding autoencoders with information theoretic concepts. arXiv: 1804.00057.

Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K., 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLOS Med. 15 (11), 1–17. doi:10.1371/journal.pmed.1002683.

Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds.), Proceedings o the ECCV. Springer International Publishing, Cham, pp. 818–833.