IJCARS manuscript No.
(will be inserted by the editor)

Positive-Gradient-Weighted Object Activation Mapping: Visual Explanation of Object Detector Towards Precise Colorectal-Polyp Localisation

Hayato Itoh · Masashi Misawa · Yuichi Mori · Shin-Ei Kudo · Masahiro Oda · Kensaku Mori

Received: date / Accepted: date

Abstract *Purpose:* Precise polyp detection and localisation are essential for colonoscopy diagnosis. Statistical machine learning with a large-scale dataset can contribute to the construction of a computer-aided diagnosis system for the prevention of overlooking and miss-localisation of a polyp in colonoscopy. We propose new visual explaining methods for a well-trained object detector, which achieves fast and accurate polyp detection with a bounding box towards a precise automated polyp localisation.

Method: We refine gradient-weighted class activation mapping for more accurate highlighting of important patterns in processing a convolutional neural network. Extending the refined mapping into multiscaled processing, we define object activation mapping that highlights important object patterns in an image for a detection task. Finally, we define polyp activation mapping to achieve precise polyp localisation by integrating adaptive local thresholding into object activation mapping. We experimentally evaluate the proposed visual explaining methods with four publicly-available databases.

H. Itoh
Graduate School of Informatics, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
Tel.: +81-52-789-5688
Fax: +81-52-789-3815
E-mail: hitoh@mori.m.is.nagoya-u.ac.jp

M. Oda and K. Mori
Graduate School of Informatics, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan

Y. Mori
Clinical Effectiveness Research Group, University of Oslo, Gaustad Sykehus, Bygg 20, Sognsvannsveien 21, Oslo, 0372, Norway

Y. Mori, M. Misawa, and S.-E. Kudo
Digestive Disease Center, Showa University Northern Yokohama Hospital, Chigasaki-chuo 35-1, Tsuzuki-ku, Yokohama, 224-8503, Japan

Results: The refined mapping visualises important patterns in each convolutional layer more accurately than the original gradient-weighted class activation mapping. The object activation mapping clearly visualises important patterns in colonoscopic images for polyp detection. The polyp activation mapping localises the detected polyps in ETIS-Larib, CVC-Clinic and Kvasir-SEG database with mean Dice scores of 0.76, 0.72 and 0.72, respectively.
Conclusions: We developed new visual explaining methods for a convolutional neural network by refining and extending gradient-weighted class activation mapping. Experimental results demonstrated the validity of the proposed methods by showing that accurate visualisation of important patterns and localisation of polyps in a colonoscopic image. The proposed visual explaining methods are useful for the interpreting and applying a trained polyp detector.

Keywords Colonoscopy · polyp detection · polyp localisation · model analysis · computer-aided diagnosis · deep learning

1 Introduction

Early detection of colorectal polyps is an essential task in colonoscopy. Especially, accurate polyp detection is indispensable since each 1% increase in adenoma detection rate was associated with a 3% decrease in interval colorectal cancer incidence [1,2]. Furthermore, a polyp's type and size are vital for colonoscopy diagnosis [3–5]. For desicions on a polyp's type and size, a precise polyp localisation is also necessary. However, there are potential risks of overlooking and incorrect localisation of polyps. Hence, a computer-aided diagnosis (CAD) system has a potential demand for the support of an endoscopist. Toward constructing a CAD system for colonoscopy, machine-learning-based polyp detectors have been proposed [1,2,6–10]. In particular, several previous works reported fast and accurate polyp-detection models of YOLO [1,2,6]. These works trained YOLO [11,12] with their large-scale in-house datasets and bounding-box annotations. Even though they achieved high polyp-detection performances, these detectors predict only rectangular bounding-box regions for polyps.

On the other hand, fully-convolutional-network (FCN)-based polyp segmentation methods such that U-Net [13] and its variants predict the pixel-wise location of polyps [7–10]. However, the construction of large-scale training data is a bottleneck for FCN-based methods since pixel-wise annotation is more complex and time-consuming than the bounding-box annotation. Currently, there is no publicly-available large dataset of colonoscopic images with pixel-wise annotations. Due to the lack of a large dataset, the previous works [7–10] adopted cross-validation by mixing several small publicly-available datasets for training and testing machine-learning methods. Since their evaluations ignore the data and annotation biases in each dataset by mixing the subsets of datasets, their cross-validation cannot validate the generalisation ability of trained models against unseen patterns. Furthermore, the previous works [8,9] performed cross-dataset evaluation by training a model with one dataset and

testing a trained model with other datasets for several public datasets. The cross-dataset evaluations showed lower evaluation values than cross-validation ones. Moreover, these datasets are designed to evaluate polyp-localisation performance, not for the training of machine-learning models. These results imply the limitation of FCN-based methods with the current public dataset to achieve a high generalisation ability.

This paper proposes new visual explaining methods for a trained object detector towards precise automated polyp localisation. Our visual explaining methods localise polyps' region by analysing a trained YOLO's activations. In other words, the proposed methods achieve polyp localisation without pixel-wise annotations of polyps. Instead of a supervised approach with FCNs, we present a practical approach for polyps' localisations. The proposed method is the first work of a precise analysis-based localisation method for a polyp detector. First, we refine Gradient-Weighted Class Activation Mapping (Grad-CAM) to highlight important patterns in a convolutional layer in processing a convolutional neural network (CNN). This refinement is achieved by selecting only positive gradients for the computation of neuron importance weights. Second, extending the refined Grad-CAM into multiscaled processing, we define Object Activation Mapping (OAM), highlighting objects' patterns in an image for YOLO's object detection. Third, by integrating local thresholding into OAM, we define Polyp Activation Mapping (PAM) to achieve precise polyp localisation. These methods highlight important patterns of an input image in the polyp-detection task step by step. Figure 1 shows the processing flow of the proposed methods. Finally, we present experimental validation of the proposed visual explaining methods by applying them to the well-trained YOLOv3 model, which achieved practical polyp-detection performance in our previous work [4]. In the experiment, we use four publicly-available databases, which are completely different from the training data of our trained YOLOv3. Thereby, we fairly evaluate the localisation performance of the proposed visualisation method.

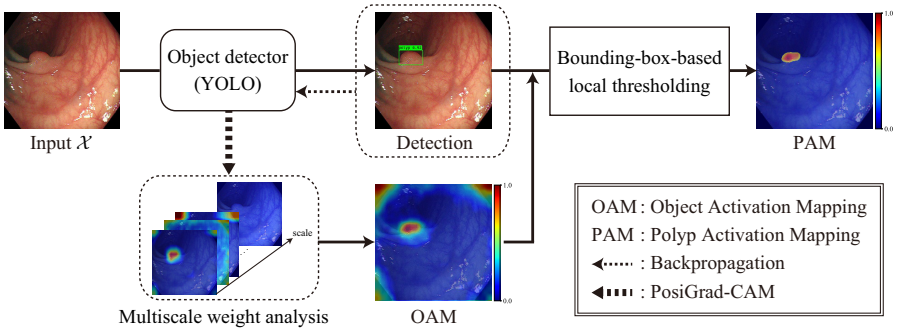


Fig. 1 Overview of our anlaysis of an object detector.

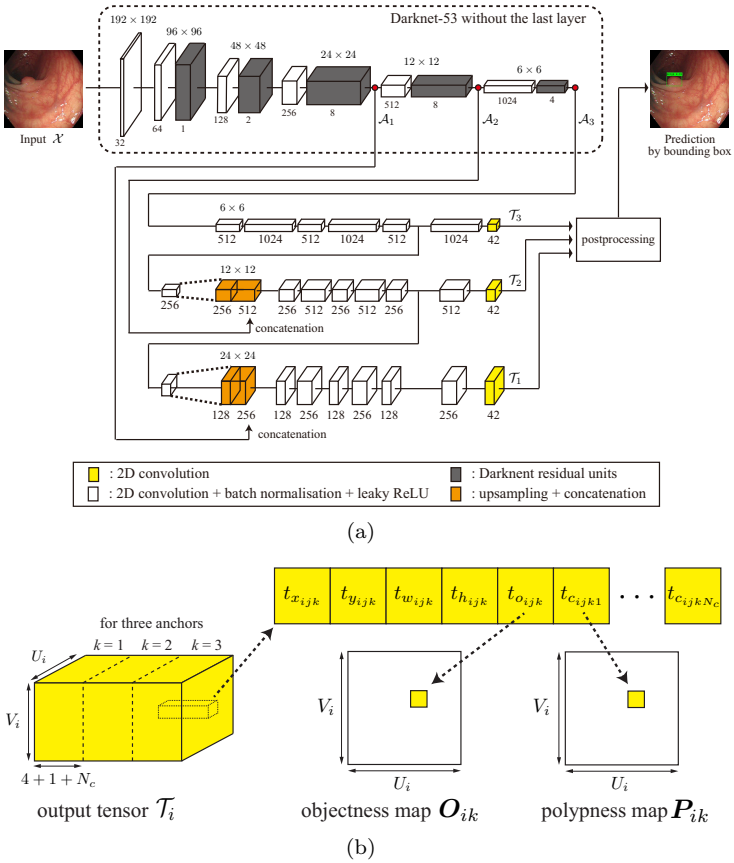


Fig. 2 Architecture and outputs of YOLOv3. (a) architecture. (b) output tensors.

2 Overview of YOLOv3

We briefly summarise the architecture, and training and test steps of YOLOv3 [12]. As shown in Fig. 2(a), YOLOv3 comprises two parts: backbone and head. The backbone part is Darknet-53 which extracts image features from input. The head part outputs predictions for three scales. In the i-th scale, YOLOv3 divides an input colour image $\mathcal{X} \in \mathbb{R}^{H \times W \times 3}$ into $U_i \times V_i$ discretised regions as cells and predicts B candidates of an object's location for each cell.

YOLOv3 uses bounding box priors. These are refered to as anchors. The k-th anchor of the i-th scale has predefined width $A_{w_{ik}}$ and height $A_{w_{ik}}$. Furthermore, YOLOv3 adopts gird-cell coordinate $(c_{x_{ij}}, c_{y_{ij}})$ in the divided $\mathcal{X}$. The origin in grid-cell coordinate is the left-top corner of $\mathcal{X}$, and $(c_{x_{ij}}, c_{y_{ij}})$ represents grid-corner indices for the j-th cell in the i-th scale. YOLOv3 computes $t_{x_{ijk}}, t_{y_{ijk}}, t_{w_{ijk}}, t_{h_{ijk}}$, and outputs a centre $(b_{x_{ijk}}, b_{y_{ijk}})$, width $b_{w_{ijk}}$, and

height $b_{h_{ijk}}$ of a location for an object by

$$b_{x_{ijk}} = \sigma(t_{x_{ijk}}) + c_{x_{ij}}, \tag{1}$$
$$b_{y_{ijk}} = \sigma(t_{y_{ijk}}) + c_{y_{ij}}, \tag{2}$$
$$b_{w_{ijk}} = A_{w_{ik}} \mathrm{e}^{t_{w_{ijk}}}, \tag{3}$$
$$b_{h_{ijk}} = A_{h_{ik}} \mathrm{e}^{t_{h_{ijk}}}, \tag{4}$$

where e is a Napier's constant and $\sigma(\cdot)$ is a sigmoid function. Furthermore, YOLOv3 computes an objectness score $\sigma(t_{o_{ijk}})$ and category scores $\sigma(t_{c_{ijk1}})$, $\sigma(t_{c_{ijk2}}), \ldots \sigma(t_{c_{ijkC}})$ for the k-th anchor of the j-th cell in the i-th scale. Therefore, for the i-th scale, YOLOv3 outputs a tensor $\mathcal{T}_i \in \mathbb{R}^{S_i \times S_i \times (B(4+1+C))}$ as shown in Fig. 2(b). By expressing all the parameters of YOLOv3 as a parameter vector $\boldsymbol{\theta}$, we define YOLOv3 as a function $f(\mathcal{X}; \boldsymbol{\theta})$ that outputs sets of objetcs' locations given by Eqs. (1)-(4), objectness scores $\sigma(t_{o_{ijk}})$, and category scores $\sigma(t_{c_{ijkl}})$ for $i = 1, 2, 3$, $j = 1, 2, \ldots, S_i \times S_i$, $k = 1, 2, \ldots, B$ and $l = 1, 2, \ldots, C$.

In the training step, YOLOv3 ignores low-confident predictions by thresholding object scores with η. For each object, YOLOv3 finds the best localisation by computing IoU (Intersection over Union) $|S \cap S^*|/|S \cup S^*|$ between a predicted region S and ground truth S^*, where $|\cdot|$ expresses the number of pixels in a region. We set weights $\mathbb{1}^{\mathrm{obj}}_{ijk} = 1$ and $\mathbb{1}^{\mathrm{obj}}_{ijk} = 0$ for the best localisation and the others, respectively, for each object. Furthermore, we set $\mathbb{1}^{\mathrm{non}}_{ijk} = 1 - \mathbb{1}^{\mathrm{obj}}_{ijk}$. By using ground truth $t^*_{x_{ijk}}$, $t^*_{y_{ijk}}$, $t^*_{w_{ijk}}$, $t^*_{h_{ijk}}$, $t^*_{c_{ijk1}}$, $t^*_{c_{ijk2}}$, $\ldots t^*_{c_{ijkC}}$ for input $\mathcal{X}$, we have a loss functional by

$$\begin{aligned}
&\mathscr{L}(f(\mathcal{X}; \boldsymbol{\theta})) = \\
&\lambda_{\mathrm{box}} \sum_{i=1}^{3} \sum_{j=1}^{S_i^2} \sum_{k=1}^{B} \mathbb{1}^{\mathrm{obj}}_{ijk} \Big(|t_{x_{ijk}} - t^*_{x_{ijk}}|^2 + |t_{y_{ijk}} - t^*_{y_{ijk}}|^2 \\
&\quad + |t_{w_{ijk}} - t^*_{w_{ijk}}|^2 + |t_{h_{ijk}} - t^*_{h_{ijk}}|^2 \Big) \\
&\quad - \sum_{i=1}^{3} \sum_{j=1}^{S_i^2} \sum_{k=1}^{B} \Big(\mathbb{1}^{\mathrm{obj}}_{ijk} \log\big(\sigma(t_{o_{ijk}})\big) \\
&\quad + \mathbb{1}^{\mathrm{non}}_{ijk} \log\big(1 - \sigma(t_{o_{ijk}})\big) \Big) \\
&\quad - \sum_{i=1}^{3} \sum_{j=1}^{S_i^2} \sum_{k=1}^{B} \sum_{l=1}^{C} \mathbb{1}^{\mathrm{obj}}_{ijk} \Big(t^*_{c_{ijkl}} \log\big(\sigma(t_{c_{ijkl}})\big) \\
&\quad + (1 - t^*_{c_{ijkl}}) \log\big(1 - \sigma(t_{c_{ijkl}})\big) \Big),
\end{aligned} \tag{5}$$

where we set $S_1 = 6, S_2 = 12, S_3 = 24$, $B = 3$, $C = 9$, and $\eta = 0.6$ as shown in Fig. 1(a). In Eq. (5), $\mathbb{1}^{\mathrm{obj}}_{ijk}$ and $\mathbb{1}^{\mathrm{non}}_{ijk}$ reduce the effects of class imbalances

between cells expressing objects and background in training and increase sensitivity and specificity in object detections. By using a training set $\{\mathcal{X}_n\}_{n=1}^N$, we search $\hat{\boldsymbol{\theta}}$ by solving

$$\arg\min_{\boldsymbol{\theta}} \mathbb{E}_n\Big[\mathscr{L}(f(\mathcal{X}_n;\boldsymbol{\theta}))\Big]. \tag{6}$$

In the testing step for a query, YOLOv3 adopts different postprocessing from the training step. For the output tensors, YOLOv3 firstly apply thresholding of scores $\zeta = \sigma(t_{o_{ijk}})\sigma(t_{c_{ijkl}})$ by τ and then applies non-maximum suppression [14,15] for detected bounding boxes, where overlapping thresholding [14] is 0.5. Finally, YOLOv3 predicts the selected bounding boxes with objectness and category scores. Note that the locations of output bounding boxes are re-scaled from grid coordinate into the coordinate of an input image.

3 Methods

3.1 Original Grad-CAM

To analyse a trained YOLOv3 model, we refine Grad-CAM [16]. For the refinement, we introduce the original definition. For a classification task, we set $y^{(c)}$ to be output of CNN for any class c. Furthermore, for $u = 1, 2, \ldots, U$, $v = 1, 2, \ldots, V$ and $m = 1, 2, \ldots, M$, we set $\boldsymbol{A}^{(m)} = (a_{uv}^{(m)})$ to be the m-th feature map of size $U \times V$ at a convolutional layer. Setting $\frac{\partial y^{(c)}}{\partial a_{uv}^{(m)}}$ a gradient of $y^{(c)}$ with respect to $a_{uv}^{(m)}$, we have a neuron importance weight

$$\alpha_m^{(c)} = \frac{1}{UV}\sum_{u=1}^{U}\sum_{v=1}^{V}\frac{\partial y^{(c)}}{\partial a_{uv}^{(m)}}. \tag{7}$$

Since $\alpha_m^{(c)}$ expresses an importance of $\boldsymbol{A}^{(m)}$ for class c [16], we can visualise important image pattern at a convolutional layer for any class c by the linear combination of feature maps

$$\boldsymbol{L}_{\mathrm{Grad}}^{(c)} = ((\ell_{uv}^{(c)})), \quad \ell_{uv}^{(c)} = \mathrm{ReLU}\left(\sum_{m=1}^{M}\alpha_m^{(c)}a_{uv}^{(m)}\right), \tag{8}$$

where $\mathrm{ReLU}(x) = \max\{x, 0\}$ for $x \in \mathbb{R}$.

3.2 PosiGrad-CAM

As guided backpropagation suppresses negative gradients [17], only positive gradients are essential for highlighting the important patterns on an image. However, the guided backpropagation leads to an over selection and results in the visualisation of textures instead of objects since a positive-gradient

neuron in a current layer can have impacts even for negative-gradient neurons in the next layer. Instead of backpropagating only positive gradients, we simply use positive gradients to compute neuron importance weights. Therefore, we redefine the definition in Eq. (7) by

$$\beta_m^{(c)} = \frac{1}{UV}\sum_{u=1}^{U}\sum_{v=1}^{V}\mathrm{ReLU}\left(\frac{\partial y^{(c)}}{\partial a_{uv}^{(m)}}\right). \tag{9}$$

Using Eq. (9), we have Positive-Gradient-Weighted Class Activation Mapping (PosiGrad-CAM) by

$$\boldsymbol{L}_{\mathrm{Posi}}^{(c)} = ((l_{uv}^{(c)})), \quad l_{uv}^{(c)} = \mathrm{ReLU}\left(\sum_{m=1}^{M}\beta_m^{(c)}a_{uv}^{(m)}\right). \tag{10}$$

3.3 Object activation mapping

To highlight important patterns of an input image in object detection, we define Object Activation Mapping (OAM) for YOLO.

For the k-th anchor in a scale $i \in \{1,2,3\}$, by re-ordering activated objectness scores $\sigma(t_{o_{ijk}})$ and l-th category scores $\sigma(t_{c_{ijkl}})$ in output tensors $\mathcal{T}_i$ over the cell index j with respect to grid-cell coordinates, we have objectness map $\boldsymbol{O}_{ik} = ((o_{uv}^{(ik)})) \in \mathbb{R}^{U_i \times V_i}$ and l-th category map $\boldsymbol{P}_{ikl} = ((p_{uv}^{(ikl)})) \in \mathbb{R}^{U_i \times V_i}$, respectively, for the k-th anchor at the i-th scale. In this paper, we refer to $\boldsymbol{P}_{ikl}$ of $l = 1$ as a polypness map $\boldsymbol{P}_{ik} = ((p_{uv}^{(ik)}))$ since we set the first categoy to be a polyp. As the same manner of the postprocessing in YOLOv3 with a criterion $\tau \in [0,1)$, we use score-based thresholding by a weight

$$\mathbb{1}_{uv}^{(ik)} = \begin{cases} 1, & \text{if} \quad o_{uv}^{(ik)}p_{uv}^{(ik)} \geq \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

In this paper, we set $\tau = 0.1$. Using the weight in Eq. (11), we define a total objectness score for the i-th scale in the head part by

$$\xi_i = \sum_{k=1}^{B}\sum_{u=1}^{U_i}\sum_{v=1}^{V_i}\mathbb{1}_{uv}^{(ik)}o_{uv}^{(ik)}. \tag{12}$$

As shown in Fig. 2, YOLOv3 bases on a multiscale feature extraction. The head part bases on the extracted feature maps of three scales, that is, tensors $\mathcal{A}_1, \mathcal{A}_2$ and $\mathcal{A}_3$ of the body part as shown in Fig. 2(a). Each tensor $\mathcal{A}_{i'}$ consists of feature maps $\boldsymbol{A}_{i'm} = ((a_{uv}^{(i'm)})) \in \mathbb{R}^{U_{i'} \times V_{i'}}$, $m = 1, 2, \ldots, M_{i'}$, where we set $M_1 = 128, M_2 = 256, M_3 = 512$, for a scale index $i' \in \{1,2,3\}$ in the body part. Note that $U_i = U_{i'}$ and $V_i = V_{i'}$. Therefore, extending Eq. (9), we define a neuron importance weight for the multiple scales by

$$\gamma_m^{(ii')} = \frac{1}{U_{i'}V_{i'}}\sum_{u=1}^{U_{i'}}\sum_{v=1}^{V_{i'}}\mathrm{ReLU}\left(\frac{\partial \xi_i}{\partial a_{uv}^{(i'm)}}\right). \tag{13}$$

By using normalised weight $\bar{\gamma}_m^{(ii')} = \frac{\gamma_m^{(ii')}}{\sqrt{\sum_{m=1}^{M_{i'}} |\gamma_m^{(ii')}|^2}}$, we have

$$\boldsymbol{L}^{(ii')} = ((l_{uv}^{(ii')})), \quad l_{uv}^{(ii')} = \mathrm{ReLU}\left(\sum_{m=1}^{M_{i'}} \bar{\gamma}_m^{(ii')} a_{uv}^{(i'm)}\right), \tag{14}$$

for scales of indexes i and i'.

After resizing $\boldsymbol{L}^{(ii')} = ((l_{uv}^{(ii')})) \in \mathbb{R}^{U_{i'} \times V_{i'}}$ into $\bar{\boldsymbol{L}}^{(ii')} = ((l_{xy}^{(ii')})) \in \mathbb{R}^{W \times H}$ of input size $W \times H$, we obtain OAM

$$\boldsymbol{L} = ((l_{xy})) = \sum_{i=1}^{3} \sum_{i'=1}^{3} w_{i'} \bar{\boldsymbol{L}}^{(ii')} \in \mathbb{R}^{W \times H}, \tag{15}$$

where we set $w_1 : w_2 : w_3 = \frac{1}{M_1} : \frac{1}{M_2} : \frac{1}{M_3}$ for balancing the difference of the number of feture maps among the three layers. This OAM highlights important patterns of an input image for object detection by using multiple output tensors and multiscale feature maps.

3.4 Polyp activation mapping

We integrate local thresholding to OAM to highlight polyp regions. YOLOv3 predicts polyps' locations by Eqs (1)-(4), and we have a mask $\boldsymbol{W} = ((w_{xy})) \in \{0,1\}^{H \times W}$, where w_{xy} expresses existence and inexistence of detected polyps by 1 and 0, respectively, at (x, y) on an input discrete image. For the detected rectangular region with the highest object score on an image, we compute a thresholding criterion κ of an OAM heatmap by Otsu's method [18]. Using $\boldsymbol{L}$, $\boldsymbol{W}$ and κ, we define Polyp Activation Mapping (PAM) by

$$\mathfrak{L} = ((\mathfrak{l}_{xy})), \quad \mathfrak{l}_{xy} = \mathrm{ReLU}\left(l_{xy} w_{xy} - \kappa\right). \tag{16}$$

4 Experiments

4.1 Settings

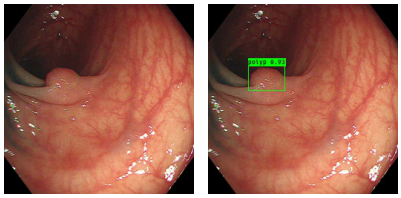
We implemented OAM and PAM, and experimentally analysed the trained YOLOv3 reported in Ref. [4] by them. We used a single GPU V100 of 32 GB (NVIDIA) and Keras with the TensorFlow backend for experiments. As described in Ref. [4], the adopted YOLOv3 was trained with 68,852 colonoscopic images collected in five hospitals with IRB approval. In training, we applied fine-tuning to the weights of the pre-trained backbone (Darknet-53 without the last layer) [12] with the stochastic gradient descent and the same data augmentations of Ref. [2]. For these images, we have only bounding-box annotations of polyps. Therefore, we used this in-house data for only training and tested the proposed methods with publicly-available databases.

We used four publicly-available databases: SUN colonoscopy video database [2], ETIS-Larib polyp [19], CVC-Clinic [20] and Kvasir-SEG databases [21]. The SUN database includes 49,136 colonoscopic images of 100 polyps and 109,554 images without a polyp. These images are extracted from colonoscopic videos as temporally successive still images. The ETIS-Larib, CVC-Clinic, and Kvasir-SEG include 196, 612, and 1000 images of polyps, respectively. These images are extracted from colonoscopic videos without temporal successiveness. In the SUN database, bounding-box annotations of polyps and pathological information, including polyps' sizes and morphologies, are available. Therefore, we used the SUN database for qualitative evaluations. On the other hand, pixel-wise annotations of polyps are available in the other three databases. We used the three databases for quantitative evaluations of localisations. Note that the trained YOLOv3 achieved AUC 0.98 and mean IoU 0.70 for the SUN colonoscopy video database, as reported in Ref. [4].

4.2 Analysi of score thresholding

To validate the object-score-based approach for the polyp localisation, we checked objectness maps $\boldsymbol{O}_{ik}$ and polypness maps $\boldsymbol{P}_{ik}$ for $i = 1, 2, 3$ and $k = 1, 2, 3$ in YOLOv3's detections. We used time-sequential colonoscopic images of the case of ID 66, where only one protruded-type polyp exists, in the SUN database as test images. Since this protruded-type polyp has a typical hemisphere shape and exists around a fold, we think this is a good example for our analysis. Figure 3 shows an example of test images and YOLOv3 output for it. As shown in Fig. 3(b), a protruded polyp is correctly detected. Figure 4 shows the visualised $\boldsymbol{O}_{ik}$ and $\boldsymbol{P}_{ik}$ for the input image in Fig. 3(a).

Figures 4(a) and (b) show objectness and polypness maps without the score thresholding of Eq. (11). In Fig. 4(a), only one cell has a non-zero value among the objectness maps. On the other hand, many cells in polypness maps have non-zero values in Fig. 4(b). This comparison clarifies that only objectness scores contribute to the localisation of a polyp, and category scores are used only for deciding whether an object is a polyp or not. After the thresholding, only one cell has a non-zero value in Figs. 4(c) and (d). Note that we confirmed



(a) (b)

Fig. 3 Examples of an input and detection result. (a) input. (b) detection result. A bounding box shows a polyp's location with its objectness.

these observations are shared characteristics of the trained YOLOv3, even for other images in ID 66 and other cases.

4.3 Qualitative evaluation of PosiGrad-CAM

To validate the advantage of PosiGrad-CAM in multiscale processing, we performed the following comparisons. First, we computed OAM heatmaps for the image in Fig. 3(a), where we used neuron importance weights of Grad-CAM and PosiGrad-CAM to compare them. In these computations, we also visualise $\bar{\boldsymbol{L}}^{(ii')}$ for $i = 1, 2, 3$ and $i' = 1, 2, 3$ of the scales in the head and body parts, respectively. Figure 5 summarises the results. Second, we computed OAM heatmaps using Grad-CAM and PosiGrad-CAM for images of 100 polyps in the SUN database. Figure 6 shows the examples of OAM heatmaps, where polyps of protruded (Is, Isp, Ip) and flat (IIa) types with different sizes (2-18 mm) exist for the two-kind computations.

Figure 5(b) shows that PosiGrad-CAM-based OAM depicted the polyp location with large values in the heatmap, while Fig. 5(a) shows that Grad-CAM-based OAM failed to capture the polyp's location. In the middle and right figures for $\mathcal{T}_1$ in Fig. 5(d), PosiGrad-CAM heatmaps have large values for the polyp's location. On the other hand, Grad-CAM heatmaps in Figs. 5(c) failed to capture the polyp's location. Furthermore, Fig. 6 clarifies that PosiGrad-CAM-based OAM captures the polyp locations for polyps of different sizes and morphologies. These comparisons clarify the validity of the PosiGrad-CAM-based computation for OAM.

4.4 Qualitative evaluation of OAM and PAM

To evaluate the temporal coherence of the proposed methods, we computed OAM and PAM heatmaps for sequential colonoscopic images in the SUN colonoscopy video database. Figure 7 shows examples of these heatmaps for still images extracted from two videos, where the protruded- and the flat-type polyps, respectively, exist.

Figure 7 shows the examples of OAM and PAM for colonoscopic videos. These results illustrate the consistent highlighting of polyps in videos. In addition to the highlight of polyps, Fig. 7(a) shows that the OAM heatmap at the furthest on the right has high values for a bubble, leading to false-positive detection. However, PAM heatmaps in the fourth row of Fig. 7(a) have high values for only polyp locations. The localisation results in the bottom row of Figs. 7(a) and (b) illustrate the shapes of protruded- and flat-type polyps. These results imply that the trained YOLOv3 is a reasonable model for the precise detection and localisation of polyps in colonoscopy.

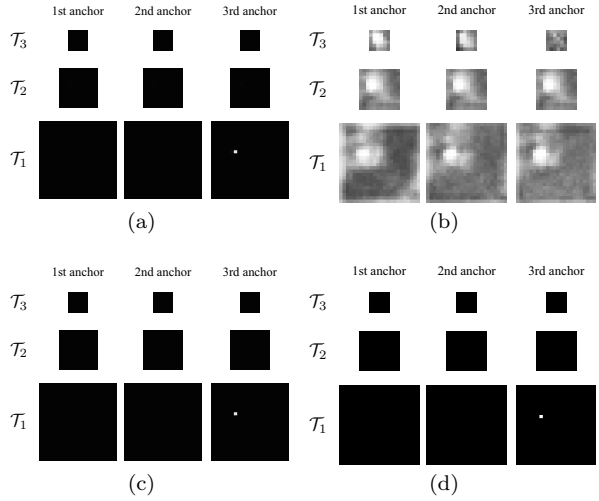


Fig. 4 Objectness and polypness maps for each anchor at three scales. The left and right columns show objectness and polypness maps, respectively. The top and bottom rows show maps for the before and after the score thresholding, respectively.

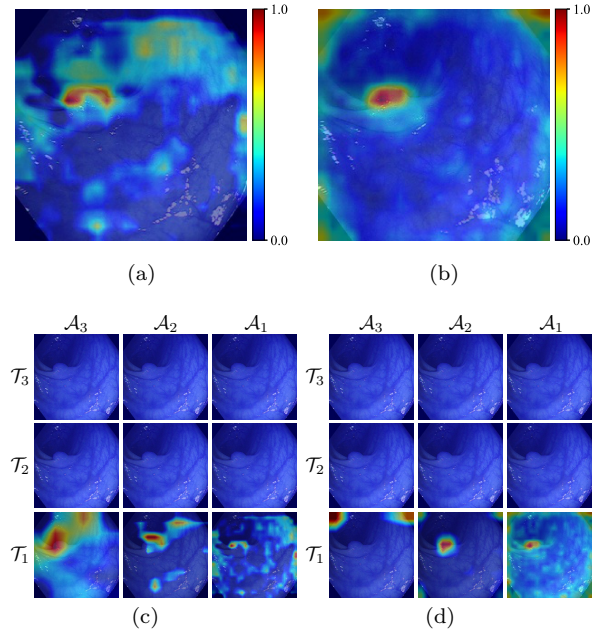


Fig. 5 Comparison of Grad-CAM and PosiGrad-CAM. (c) and (d) show the heatmaps of Grad-CAM and PosiGrad-CAM computation, respectively. In (c) and (d), row and column express head-part scale i and body-part scale i', respectively. (a) and (b) show OAM heatmaps based on Grad-CAM and PosiGrad-CAM computation, respectively. In (a), we set $w_{i'} = 1$ for $i' = 1, 2, 3$.

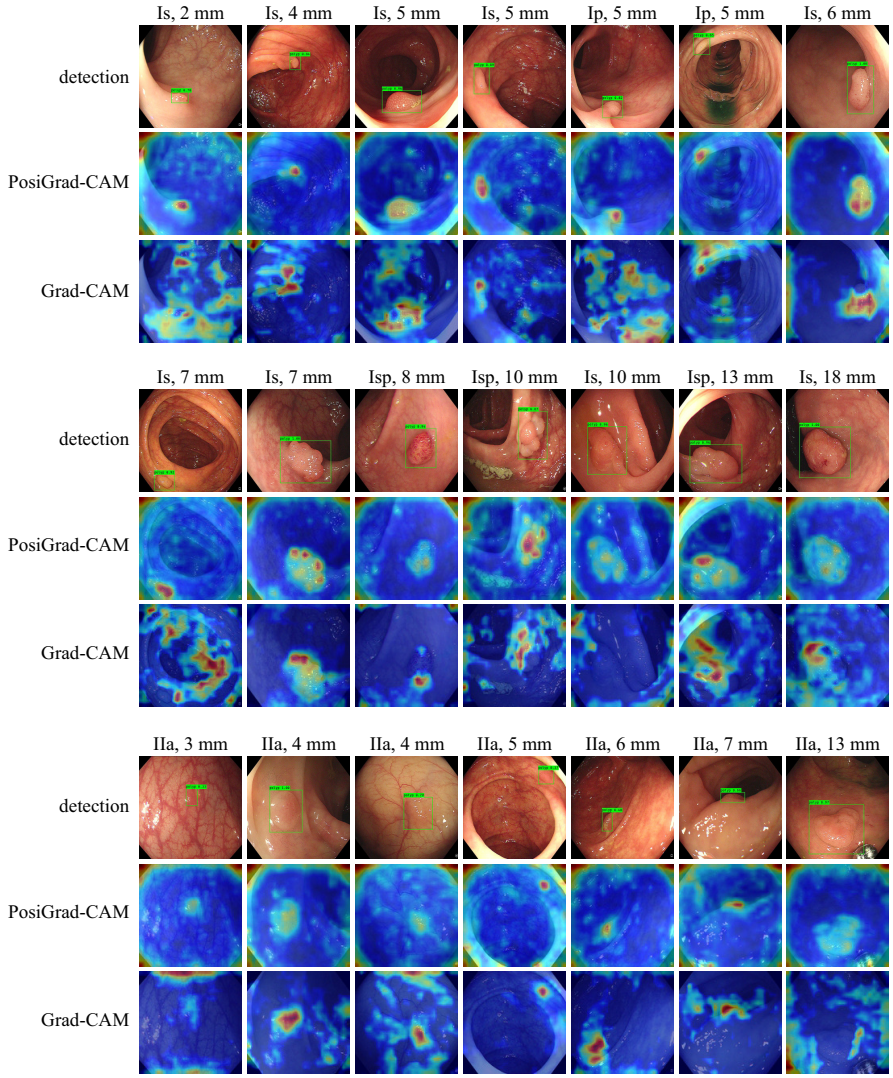


Fig. 6 Comparison of OAM heatmaps between PosiGrad-CAM and Grad-CAM based computations.

4.5 Quantitative evaluation of PAM

Finally, we quantitatively evaluated the localisation accuracy of PAM by using three databases: ETIS-Larib, CVC-Clinic and Kvaseir-SEG databases. We set a non-zero region in a PAM heatmap as an estimated polyps' region R. For a ground-truth R^* and R, we used a Dice score $(2|R \cap R^*|)/(|R| + |R^*|)$ as an evaluation value, where $|\cdot|$ expresses the number of pixels in a region.

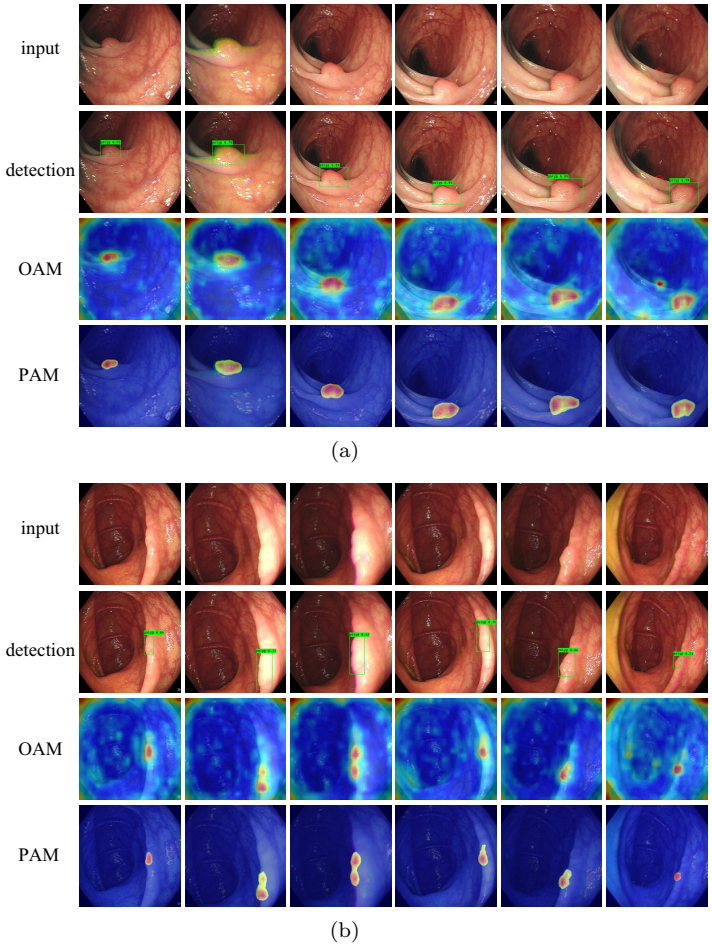


Fig. 7 Examples of the proposed visual explanation for colonoscopic videos. In (a) and (b), images from the top to bottom rows show input images, detection results, OAM heatmaps and PAM heatmaps respectively.

Furthermore, we defined detection rate as the ratio of detected images in a database, where each detected image has a Dice score equal to or larger than δ. Figure 8 summarises the detection rates and mean Dice scores for the three databases. Figure 9 shows the examples of localisation results for the three databases.

Since the proposed method is the first work of the analysis-based object localisation for an object detector, a direct comparison of performances among the proposed and state-of-the-art methods is unavailable. Instead of the direct comparison, we presented the two-kind comparisons. As the first comparison, we compared the localisation accuracy of PAM between Grad-CAM and

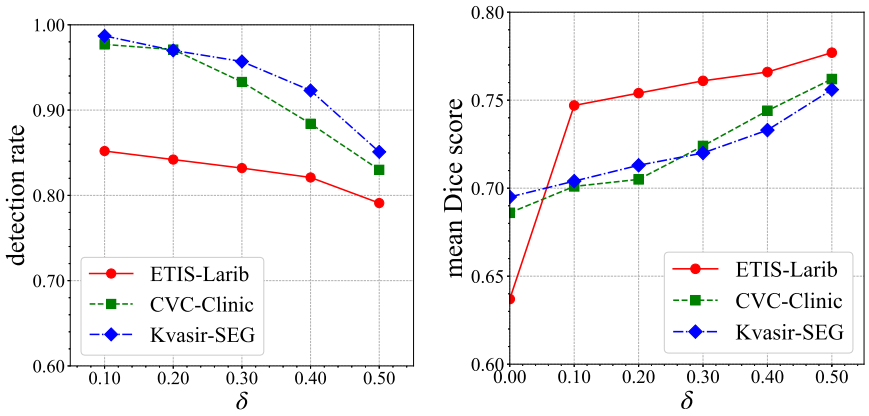


Fig. 8 Evaluation of PAM-based detection and localisation. (a) detection rate. (b) mean Dice score. In (a) and (b), δ expresses a thresholding criteria.

Table 1 Comparison of localisation accuracy of PAM between Grad-CAM and PosiGrad-CAM based computations with $\delta = 0.1$. We used a mean Dice score as a metric.

	PosiGrad-CAM	Grad-CAM
CVC-Clinic	0.701	0.423
ETIS-Larib	0.747	0.406
Kvasir-SEG	0.704	0.452

Table 2 Cross-dataset evaluations of FCN-based methods reported in previous works [8,9]. The Dice scores of ResNet++ with conditional random field and test-time argumentation, and the ones of U-Net, DoubleU-Net, BA-Net and PolypSegNet are reported in Ref. [8] and Ref. [9], respectively.

Training	Testing	U-Net [13]	ResUNet++ [8]	Double U-Net [7]	BA-Net [10]	Polyp SegNet[9]
Kvasir-SEG	CVC-Clinic	0.750	0.671	0.753	0.766	0.781
	ETIS-Lalib	0.602	0.400	0.644	0.671	0.718
CVC-Clinic	Kvasir-SEG	0.668	0.721	0.676	0.684	0.702
	ETIS-Lalib	0.575	0.397	0.612	0.637	0.686

PosiGrad-CAM based computation with $\delta = 0.1$ as shown in Table 1. For the second comparison, we quoted the state-of-the-art performances of FCNs in cross-datasets evaluations [8,9], as shown in Table 2, since we don not have pixel-wise annotations of polyps for our in-house data.

Figure 8 shows the validity of PAM. In Fig. 8(a), the detection rates with $\delta = 0.1$ for ETIS-Larib, CVC-Clinic and Kavasir-SEG are 0.852, 0.977 and 0.987, respectively. Furthermore, the mean Dice scores of correctly detected polyps for $\delta = 0.10, 0.20, \ldots, 0.50$ in the three databases are over 0.70 in Fig. 8(b). These results clarify the generalisation ability of our trained YOLOv3 and PAM both for polyp detection and localisation. Furthermore, Table 1

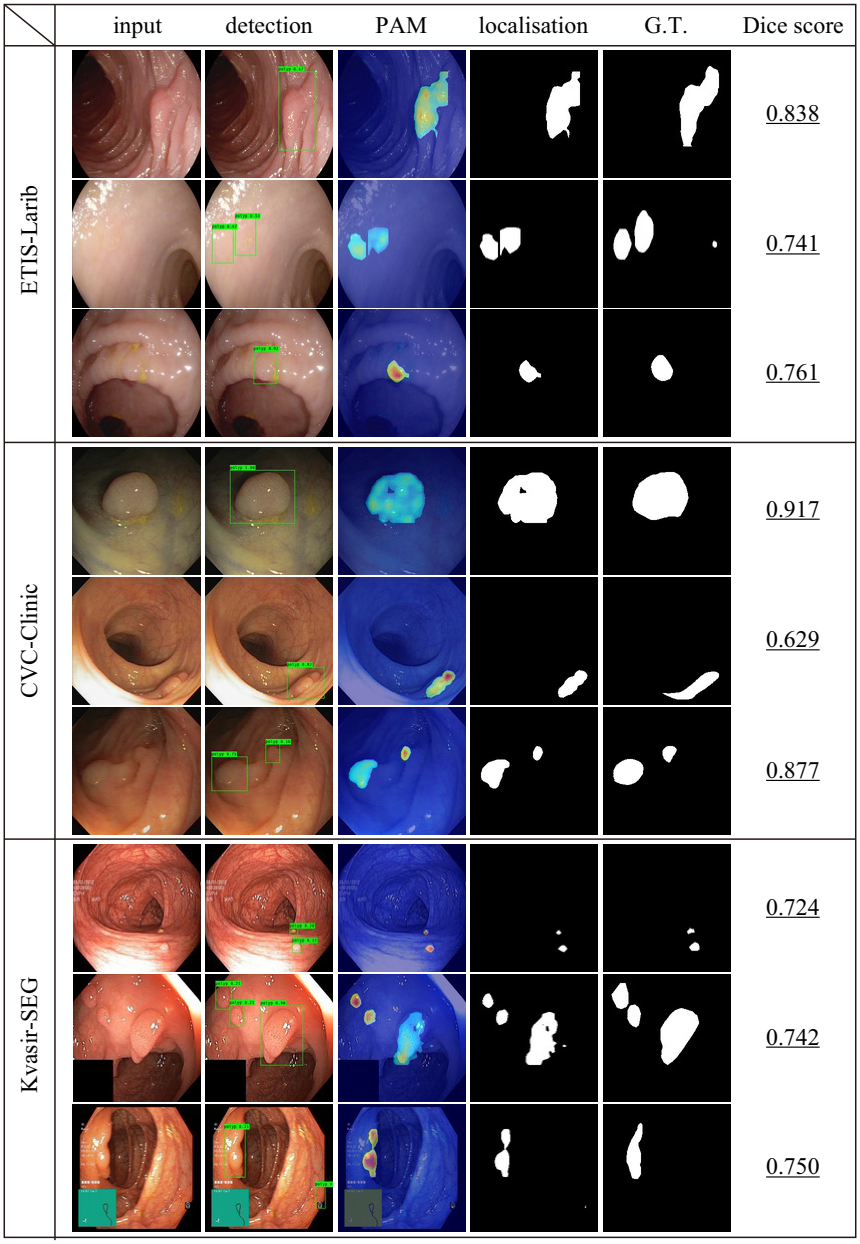


Fig. 9 Examples of PAM-based localisation in three databases: ETIS-Larib, CVC-Clinic and Kvasir-SEG databases.

shows that the PAM computed by PosiGrad-CAM outperformed the one computed by the original Grad-CAM. Moreover, the comparison between Fig. 8(b) and Table 2 shows the higher generalisation ability of PAM than FCN-based methods.

5 Discussion

Experimental results in the analysis of score thresholding and qualitative evaluation of PosiGrad-CAM demonstrated the validity of utilising multiscale object scores, selecting positive gradients, and processing multiscale feature maps for the visual explanation. Even though Grad-CAM generally visualises blurred and corrupted shapes of objects only for a simple classification task, the proposed OAM accurately visualise important object patterns in the detection of polyps with different sizes (2-18 mm) and morphologies (Is, Isp, Ip, and IIa).

Qualitative evaluations of OAM and PAM show their stable performance against temporal changes in each colonoscopic video. This temporal coherence of the visual explanations is necessary for practical applications since unstable performance is unconvincing for endoscopists. Furthermore, comparing the visualisations between OAM and PAM clarifies the patterns a detector reacts and selects in processing. For example, in Fig. 7(a), PAM visualises only a detected polyp, whilst OAM visualises injected water and a bubble and polyp. From these visualisations, we can examine which patterns are essential for a detector's processing. We think our visual explaining methods offer clues for interpreting the detection results.

Quantitative evaluations of PAM with the three databases demonstrate the advantage of PAM towards practical application to colonoscopy. In Fig. 8(a), the polyp detection with the PAM-based rejection achieved high detection rates in the three databases. In Fig. 8(b), PAM worked well in polyp localisations for all the three databases with $\delta \geq 0.1$. As shown in Fig. 9, the primal polyps' locations are correctly computed, while the regions around ambiguous boundaries between colon walls and polyps result in wrong localisations. However, annotators' biases might exist in the G.T. labels since a boundary between a colon wall and polyp is essentially ambiguous. Even for the three databases with different annotators' biases, that is, G.T. labels including uncertainty, our PAM averagely works well. These results show that the proposed method localises a polyp accurately if a given model detects a polyp. Furthermore, Tables 1 and 2 also support the validity of PAM. In Table 1, our PosiGrad-CAM based PAM outperformed Grad-CAM based one. Comparing scores in Table 2 and Fig. 8(b), PAM achieved more stable localisation for all the databases than the state-of-the-art methods trained with the publicly-available databases.

6 Conclusions

This paper proposed visual explaining methods for precise polyp localisation. The series of experimental evaluations demonstrated the validity of the methods for polyp localisation. In addition to PAM's localisation of detected polyps, OAM's visualisation of important object patterns increases the interpretability of detection results. Even though the existing FCN-based methods with publicly-available databases achieved insufficient generalisation ability of polyp localisation, our visual explaining method localise polyps accurately even in unseen data if the trained YOLO has a generalisation ability for the detection task.

Acknowledgements This study was funded by grants from AMED (19hs0110006h0003), JSPS MEXT KAKENHI (26108006, 17H00867, 17K20099), and the JSPS Bilateral Joint Research Project.

Conflicts of interest

Kudo SE recieved scholaship grant from TAIHO Pharmaceutical Co. Ltd., CHUGAI Pharmaceutical Co. Ltd. and Bayer Yakuhin Ltd. Misawa M received lecture fees from Olympus. Mori Y received consultant and lecture fees from Olympus. Mori K is supported by Cybernet Systems and Olympus (research grant) in this work and by NTT outside of the submitted work. The other authors have no conflicts of interest.

Ethical approval

All the procedures performed in studies involving human participants were in accordance with the ethical committee of Nagoya University (No. 357), and the 1964 Helsinki declaration and subsequent amendments or comparable ethical standards. Informed consent was obtained by an opt-out procedure from all individual participants in this study.

References

1. 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% Accuracy in Screening Colonoscopy. Gastroenterology 155(4): 1069-1078.e8
2. Misawa M, Kudo S-E, Mori Y, Hotta K, Ohtsuka K, Matsuda T, Saito S, Kudo T, BaBa T, Ishida F, Itoh H, Oda M, Mori K (2021) Development of a Computer-Aided Detection System for Colonoscopy and a Publicly Accessible Large Colonoscopy Video Database (with video). Gastrointestinal Endoscopy 93(4): 960-967
3. Lieberman DA, Rex DK, Winawer SJ, Giardiello FM, Johnson DA, Levin TR (2012) Guidelines for Colonoscopy Surveillance after Screening and Polypectomy: a Consensus Update by the US Multi-Society Task Force

4. Itoh H, Oda H, Jiang K, Mori Y, Misawa M, Kudo SE, Imai K, Ito S, Hotta K, Mori K (2021) Binary Polyp-size Classification based on Deep-Learned Spatial Information. International Journal of Computer Assisted Radiology and Surgery 16: 1817-1828
5. Itoh H, Oda H, Jiang K, Mori Y, Misawa M, Kudo SE, Imai K, Ito S, Hotta K, Mori K (2021) Uncertainty Meets 3D-Spatial Feature in Colonoscopic Polyp-Size Determination. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, in Print.
6. Nogueira-Rodríguez A, Domínguez-Carbajales R, Campos-Tato F, Herrero J, Puga M, Remedios D, Sánchez E, Iglesias Á, Cubiella J, Fdez-Riverola F, López-Fernández H, Reboiro-Jato M, Glez-Peña D (2021) Real-Time Polyp Detection Model using Convolutional Neural Networks. Neural Computing and Appliccations, In Print
7. Jha D, Riegler MA, Johansen D, Halvorsen P, Johansen HD (2020) DoubleU-Net: A Deep Convolutional Neural Network for Medical Image Segmentation. Proc. IEEE 33rd International Symposium on Computer-Based Medical Systems: 558-564
8. Jha D, Smedsrud PH, Johansen D, de Lange T, Johansen HD, Halvorsen H, Riegler MA (2021) A Comprehensive Study on Colorectal Polyp Segmentation With ResUNet++, Conditional Random Field and Test-Time Augmentation. IEEE Journal of Biomedical and Health Informatics 25(6): 2029-2040
9. Mahmud T, Paul B, Fattah SA (2021) PolypSegNet: A Modified Encoder-Decoder Architecture for Automated Polyp Segmentation from Colonoscopy Images. Computers in Biology and Medicine 128: 104119
10. Wang R, Chen S, Ji C, Fan J, Li Y (2022) Boundary-Aware Context Neural Network for Medical Image Segmentation. Medical Image Analysis 78: in Print.
11. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You Only Look Once: Unified, Real-Time Object Detection. Proc. IEEE International Conference on Computer Vision: 779-788
12. Redmon J, Farhadi A (2018) YOLOv3: An Incremental Improvement. CoRR arXiv:1804.02767v1, `https://pjreddie.com/darknet/yolo/`
13. Ronneberger O, Fischer P, Brox T, Navab N, Hornegger J, Wells WM, Frangi A (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation, Proc. Medical Image Computing and Computer-Assisted Intervention LNCS 9351: 234–241
14. Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. Proc. Neural Information Processing Systems: 91-99
15. Bodla N, Singh B, Chellappa R, Davis LS (2017) Soft-NMS Improving Object Detection with One Line of Code. Proc. IEEE International Conference of Computer Vision: 5562-5570
16. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2020) Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision 128: 336-359
17. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller MA (2014) Striving for Simplicity: The All Convolutional Net. Proc. ICLR workshop trak (2015)
18. Otsu N (1979) A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems, Man, and Cybernetics 9(1): 62-66
19. Silva JS, Histace A, Romain O, Dray X, Granado B (2014) Towards Embedded Detection of Polyps in WCE Images for Early Diagnosis of Colorectal Cancer International Journal of Computer Assisted Radiologyand Surgery 9(2): 283-293
20. Bernal J, Sánchez FJ, 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. Computerized Medical Imaging and Graphics 43: 99-111
21. Jha D, Smedsrud PH, Riegler MA, Halvorsen P, Johansen D, de Lange T, Johansen HD (2020) Kvasir-SEG: A Segmented Polyp Dataset, Proc. International Conference on Multimedia Modeling