Information Fusion 92 (2023) 64–79

Contents lists available at ScienceDirect

Information Fusion

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

# Fusion of colour contrasted images for early detection of oesophageal squamous cell dysplasia from endoscopic videos in real time

Xiaohong W. Gao [a,*,2], Stephen Taylor [b], Wei Pang [c], Rui Hui [e], Xin Lu [d], Oxford GI Investigators[f,1], Barbara Braden [f,*,2]

[a] Middlesex University, London, United Kingdom
[b] University of Oxford, Oxford, United Kingdom
[c] Heriot-Watt University, Edinburgh, United Kingdom
[d] Ludwig Institute, University of Oxford, Oxford, United Kingdom
[e] Beijing 301 Hospital, China
[f] Translational Gastroenterology Unit, John Radcliff Hospital, University of Oxford, Oxford, United Kingdom

ARTICLE INFO

*Keywords:*
Early squamous cell cancer detection
Deep machine learning
Colour contrast enhancement
Endoscopic treatment
Surveillance
Oesophagus cancer
gastrointestinal endoscopy

ABSTRACT

Standard white light (WL) endoscopy often misses precancerous oesophageal changes due to their only subtle differences to the surrounding normal mucosa. While deep learning (DL) based decision support systems benefit to a large extent, they face two challenges, which are limited annotated data sets and insufficient generalisation. This paper aims to fuse a DL system with human perception by exploiting computational enhancement of colour contrast. Instead of employing conventional data augmentation techniques by alternating RGB values of an image, this study employs a human colour appearance model, CIECAM, to enhance the colours of an image. When testing on a frame of endoscopic videos, the developed system firstly generates its contrast-enhanced image, then processes both original and enhanced images one after another to create initial segmentation masks. Finally, fusion takes place on the assembled list of masks obtained from both images to determine the finishing bounding boxes, segments and class labels that are rendered on the original video frame, through the application of non-maxima suppression technique (NMS). This deep learning system is built upon real-time instance segmentation network Yolact. In comparison with the same system without fusion, the sensitivity and specificity for detecting early stage of oesophagus cancer, i.e. low-grade dysplasia (LGD) increased from 75% and 88% to 83% and 97%, respectively. The video processing/play back speed is 33.46 frames per second. The main contribution includes alleviation of data source dependency of existing deep learning systems and the fusion of human perception for data augmentation.

## 1. Introduction

This paper introduces the fusion of colour contrasted images and their original counterparts to improve the detection accuracy for early diagnosis of oesophageal cancer and precancerous changes during endoscopic procedures in real time.

Oesophagus cancer (EC) remains the 9th most common cancer [1] and the 6th leading cause of cancer-related death [2] in the world. In 2018, the estimated number of new cases was 572,000, of which approximately 509,000 persons (89%) died from oesophageal cancer [1]. Histologically, there are two major types that constitute the majority of all oesophageal cancers, adenocarcinoma and squamous cell carcinoma cancer (SCC) (87%) [3,4].

While the overall five-year survival rate of oesophagus cancer is less than 20% [5], this figure can be improved significantly to up to 90% if an oesophageal cancer is detected in its intramucosal stage when lymph node metastasis is unlikely, and endoscopic resection or surgery is possible. As reported by Naito et al. [6] and Takana et al. [7], endoscopic

* Corresponding authors.
*E-mail addresses:* x.gao@mdx.ac.uk (X.W. Gao), braden@em.uni-frankfurt.de (B. Braden).
[1] Oxford GI Investigators for data contribution and curation: Philip Allan, Tim Ambrose, Carolina Arancibia-Cárcamo;, Adam Bailey, Ellie Barnes, Elizabeth Bird-Lieberman, Jan Bornschein, Oliver Brain, Jane Collier, Emma Culver, James East, Alessandra Geremia, Bruce George, Lucy Howarth, Kelsey Jones, Paul Klenerman, Simon Leedham, Rebecca Palmer, Fiona Powrie, Astor Rodrigues, Jack Satsangi, Alison Simmons, Simon Travis, Holm Uhlig, Alissa Walsh.
[2] Both Gao and Braden are corresponding authors.

https://doi.org/10.1016/j.inffus.2022.11.023
Received 9 September 2022; Received in revised form 14 November 2022; Accepted 20 November 2022
Available online 24 November 2022

or surgical resection can achieve excellent curative outcome as long as the oesophageal cancer is confined to the first layer of the oesophageal wall, i.e. intramucosal stage (T1a), and the tumour has not gone beyond the first layer of the oesophageal wall as then lymph node metastasis is highly unlikely. However, if the tumour reaches the second layer, the submucosal stage (T1b), the risk of lymph node invasion is already substantial. Overall, the 5-year survival rate for T1a patients is 94% and 72% for T1b patients.

Unfortunately, routine upper gastrointestinal endoscopy carries a significant miss rate for detecting oesophageal cancer and precancerous lesions due to their inconspicuous changes in the surface appearance in the early intra-mucosal stage, which is determined by a number of research groups, including Georgina et al. [8], Chai et al. [9], and de Santigo at al. [10]. As a result, around 25% [11], i.e. 1 in 4, of patients of oesophageal cancer were given normal findings the year before when the diseased regions only presented subtle changes in comparison with normal mucosa (oesophageal linig).

The challenges clinicians face in detection of precancerous changes in squamous epithelium and early stages of SCC are the inconspicuous appearance of affected regions and detection speed. To minimise patients' discomfort while undergoing endoscopy, time is limited, usually ~10mins are scheduled, assigning clinicians to inspect ~18,000 frames (=30 frames/*sec* × 60 s × 10 min) in such short period. Furthermore, the early onset of SCC grows usually flat with only subtle changes in appearance in both colour and microvasculature compared to normal epithelium when the endoscopy is performed as conventional white light endoscopy (WLE). Fig. 1 exemplifies some neoplastic lesions in squamous epithelium where red colour refers to 'cancer ', blue to 'High grade of dysplasia' (HGD) and green to 'Low grade dysplasia' (LGD). The suspicious regions are delineated by clinicians and had been histologically confirmed by targeted biopsies.

The response of tissues to an illuminating endoscopic light strongly depends on the tissue properties and on the spectrum the light accommodates. Under conventional WLE (Fig. 1(a)) pre-cancerous squamous neoplasia present discrete variations with subtle changes to normal tissue. While narrow-band imaging (NBI) [12] (Fig. 1(b)) takes advantage of spectral principles by employing two wavelengths at 415*nm* (blue) and 540*nm* (green), it is confined to only two mono-colour bands. Another imaging approach is dye-based chromo-endoscopy, e.g. Lugol's staining technique, which highlights dysplastic abnormalities with depleted glycogen storages by spraying iodine [13], producing images with orange-like colours and unstained areas of dysplasia (Fig. 1(c).

NBI technique mainly facilitates the detection of unique vascular pattern morphology that are present in neoplastic lesions [14]. However, precancerous stages can take a variety of forms which sometimes are difficult to recognise (Fig. 1(b) arrow). On the other hand, for Lugol's staining approach, some patients react uncomfortably to the iodine spray, which limits its application.

Hence, it is of a clinical priority to have a computer assisted diagnostic (CAD) system that supports clinicians' decision-making in real time by highlighting potentially neoplastic regions while patients are undergoing endoscopic inspection. In this way, the system can prompt to take a biopsy from the correct spot, which will lead to facilitating endoscopic treatment by delineating the lesion and identifying patients in need for surveillance, all to prevent progression to cancer.

For the development of such CAD systems applying deep machine learning techniques, the main obstacles encountered are the lack of labelled ground truth dataset and insufficiency of generalisation, especially in the medical domain.

To overcome these hurdles, many researchers capitalise on a number of well-known techniques to enhance system robustness and performance. These techniques include transfer learning to apply pre-trained networks instead of training from scratch, weakly/unsupervised learning to analyse images only with limited labelling, generative frameworks to learn to generate images allowing algorithms understand main distinctive features and multitask learning to learn interrelated concepts in an attempt to produce better generalisations [15,16].

To address data insufficiency, several research teams work on the findings of the optimal number of dataset in order to achieve the best system performance [17,18]. While a larger number of data contributes to higher accuracy in results, it appears that the performance reaches a plateau at a certain data size point. This tends to be domain orientated. In addition, the performance of a developed system is dependant on its ability for generalisation [19,20].

In deep leaning community, *Generalisation* remains one of the fundamental unsolved problems. A model optimised on a finite set of training data usually does not perform well on a held-out test set [21]. This is because there is gap between theory and practice. This gap is exacerbated when a model is over-parameterised, by which the theory has the capacity to overfit the given train sets but often does not in practice. One solution, as proposed by Nakkiran et al. [21], is to perform online optimization to allow the trained model to access to an infinite stream of sample data and hence to update and adapt iteratively and constantly. This approach, however, has challenges when it is applied to the medical domain where confidential patient data should not be distributed nor placed online. Universal consenting for anonymised data feeding might allow this in medicine.

In Endoscopy, a number of different approaches have been proposed to generalise the trained models by augmenting datasets in various forms, for example, further enhancing and highlighting neoplastic

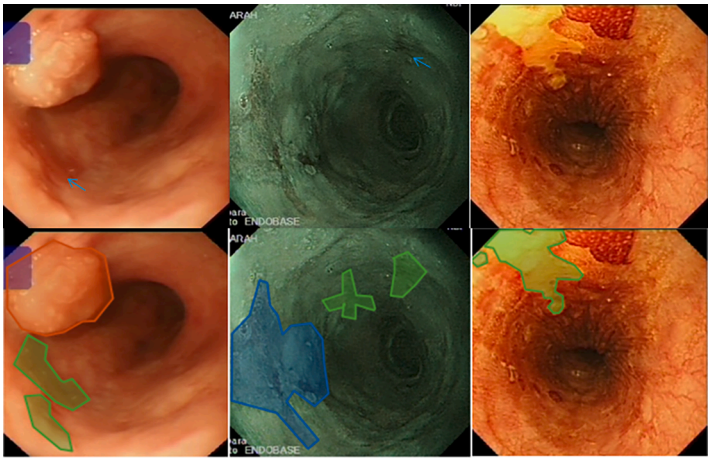


**Fig. 1.** Examples of diseased regions where boundary red='cancer', green='LGD', blue='HGD'. Top row: original images; bottom: with labelled masks delineated by the experts. (a) WLE; (b) NBI; (c) Lugol's. Arrows pointing to the lesioned regions with subtle changes in comparison with surrounding normal tissue.

regions. While spectral [22] or multi-spectral imaging (with 4 to 16 mono spectrum bands) or hyper-spectral imaging [23] (16 to 40 bands) techniques have demonstrated potential to depict endogenous contrast capitalising on wavelength-dependant light-tissue interactions [24,25], these systems require complex designated optical devices to acquire spectral signals, giving rise to operational difficulties, including physical implementation, prolonged acquisition time and post co-registration.

Hence, this study proposes a novel approach to improve endoscopic detection of pre-cancerous changes in squamous epithelium while leveraging the issue of data shortage. It fuses the enhanced images with the original ones by increasing the colour contrast of neoplastic areas to their surroundings. To alleviate the insufficiency of generalisation of the developed system, this contrast enhancement originates from human colour perception by applying the colour appearance model of CIECAM02 [26], standardised by the *Commission Internationale de l'éclairage* (CIE). CIECAM02 and its recently simplified version CAM16 [27] are modelled to simulate human visual perception to transform between physical colour spectral values of CIE tristimulus (XYZ) and perceptual attribute correlates, lightness (*J*), colourfulness (*C*) and Hue (*H*), by taking the viewing conditions into account. These contrast-modified endoscopic images are then added to the training of a deep learning-based system for detection, delineation and classification of LGD, HGD or SCC. Upon testing, for each frame, the system will generate its contrasted one automatically and process them together with the final fused detection results displayed on the original image.

The remaining of this paper is structured as follows. Section 2 briefly reviews the state-of-the-art deep learning architectures for performing real time tasks of detection, segmentation as well as classification, which is followed by Section 3 that entails the methodology employed in this work. The results are presented in Section 4, which leads to the discussion and conclusion in Sections 5 and 6 respectively.

## 2. Related work

### 2.1. State of the art deep learning systems for analysis of oesophageal images and videos

Progress on diagnosis of oesophageal cancer through the application of artificial intelligence (AI) using convolutional neural networks (CNN) has been made by several research teams recently [28,29]. For example, research conducted by Horie et al. [30], distinguishes oesophageal cancers from non-cancer patients with an aim to evaluate diagnostic accuracy. While applying conventional CNN architecture to classify two classes, they are able to achieve 98% sensitivity for cancer detection. The work carried out by Ghatwary et al. [31], evaluate several state of the art (SOTA) CNN approaches, aiming to achieve early detection of SCC from high-definition white light endoscopy (HD-WLE) images, and conclude that the approaches of SSD [32] and Faster R-CNN [33] perform better. Again, two classes are investigated in their study, i.e. cancer and normal subjects. While these studies exhibit high accuracy of classification, the main focus of those research remains on the binary classification of normal from abnormal. With regarding to early detection of any potential suspicious regions regardless how small they are, segmentation of abnormal regions also plays a key role in delegating clinical decisions. This is because the collection of a biopsy, as well as treatment, necessitates to pin point the exact spot while clinicians also are negotiating with the movements of the heart, respiration, peristalsis and endoscopic camera during endoscopy procedures.

In addition, in order to assist clinicians with the diagnosis while performing endoscopy, real-time processing of videos, i.e. with processing speed of 24+ frames per second (fps) or at most 41 milliseconds (ms) per frame, should be realised. The work carried out by Everson et al. [34], is able to achieve inference time between 26 and 37 (ms) for an image while attempting to perform characterisation of abnormities by applying AI techniques. However, their image size appears to be half of ours at a resolution of 696 × 308 pixels. More recently, the decision-making support system by Guo et al. [35], can realise video processing times at 25 frames per second, which however is only applied for narrow band images (NBI). Table 1 summarises the current development in assisting diagnosis of oesophageal cancers.

As addressed above, these existing studies focus mainly on binary classification of endoscopic images between normal and grossly abnormal stages with little work providing bounding boxes of suspicious regions (detection) and delineation (segmentation), which is especially important when an image contains multiple lesions of varying diseased grades.

To segment an image, there are approaches of two-stage and one-stage. The region-based CNN, or R-CNN [39] family comprises two major steps. The first step proposes a set of regions of interests by selective search. Then a classifier e.g. *Support Vector Machine* (SVM), is applied to process those candidate regions in this second step.

### 2.2. Real-time processing

While these region-based object detection algorithms can achieve high accuracy, they are too slow for real-time video processing at about ~1 second per frame [38] while applying masque R-CNN [40]. Hence single-stage approaches are sought after.

One-stage method skips the region proposal stage and runs detection directly over a dense sampling of possible locations. As a result, this approach is faster and simpler, but might potentially pull down the performance to a certain extent. This one-shot category includes models of SSD [32], YOLO family [41] and RetinaNet [42]. In comparison, RetinaNet performs the best in accuracy whereas YOLOv3 runs 3.8x faster and achieves better and faster results than SSD.

Although all these one-stage approaches can achieve better performance with fast processing speed, they don't provide masks, i.e., segmentation, of the objects in concern, which limits their applications to a certain extent. This is because taking a biopsy requires a precisely defined/segmented location. From computation point of view, yielding masks in additional of bounding boxes inevitably increases processing time, which hampers the development of real time systems.

More recently, the network of Yolact [43] (you only look at coefficient) that is built upon one-stage RetinaNet by adding a masque branch, not only can provide instance segmentation but also is able to achieve real-time inference with an average 33.5 frames per second (fps) on MS COCO datasets. In this study, Yolact is applied with the fusion of contrasted images.

## 3. Methodology

### 3.1. The architecture of fused system for real time processing of endoscopic videos

Fig. 2 outlines the architecture of the fused system developed in this study. When an image (2(a)) is loaded, its contrasted counterpart (2(b)) is generated by the system (to be elaborated in 3.2). After producing bounding box regression coefficients and class confidences for each of original (2(d)) and contrasted images (2(e)) applying Yolact model (2(c)) [43], the fusion of final detection results takes place (2(f)). All the detected anchors/regions from both images are assembled together with the final determined detectors being manifested on the original image. As such, the non-maxima suppression technique (NMS) [40,43] is employed to determine whether an instance should be kept or discarded. The duplicated detections are suppressed not only for each class, but also for cross-class boxes. For example, in Fig. 2, the probability for a region to be a 'cancer' is 0.64 (2(d)) whereas the same region detected on the contrasted image has a likelihood of 0.99 to be 'low grade' (2(e)). Hence the classification outcome for this concerned region is ascertained as LGD (2(f)).

In Fig. 2, for training, the enhanced images (2(b)) are generated in advance and fed into the deep learning network of Yolact (2(c)) together

**Table 1**
Summarisation of current SOTA systems on endoscopy.

| Refs. | Modality | Class (Number) | Approach | Image Size (max pixel) | Sensitivity (%) | Specificity (%) | Speed (fps) |
|---|---|---|---|---|---|---|---|
| Ohmori et al. [28] | WLE, NBI | 2 | CNN | 300 × 300 | 100 | 69 | |
| De Groof et al. [29] | Barrett | 2 | ResNet-UNet | 256 × 256 | 93 | 83 | |
| Horie et al. [30] | WLE, BNI | 2 | CNN | 300 × 300 | 98 | 79 | 41.1 |
| Ghatwary et al. [31] | HD-WLE | 2 | R-CNN | 512 × 512 | 96 | 92 | |
| Everson et al. [34] | ME-NBI | 4 | CNN | 696 × 308 | 89.7 | 96.9 | 27 |
| Guo et al. [35] | NBI | 2 | CNN | 34mm | 98.0 | 95.0 | 25 |
| Mashimo et al. [36] | VCE | 2 | CNN | 1024 × 1024 | 98 | | 41.1 |
| Tsai et al. [23] | Hyperspectral | 3/4 | VGG | | 91 | 94 | |
| Dumoulin et al. [37] | WL (Barrett) | 2 | CNN | | 96 | 92 | |
| Gao et al. [38] | NBI, WLE, Lugol's | 3 | YoloV3 | 1920 × 1080 | 84 | 89 | 15 |

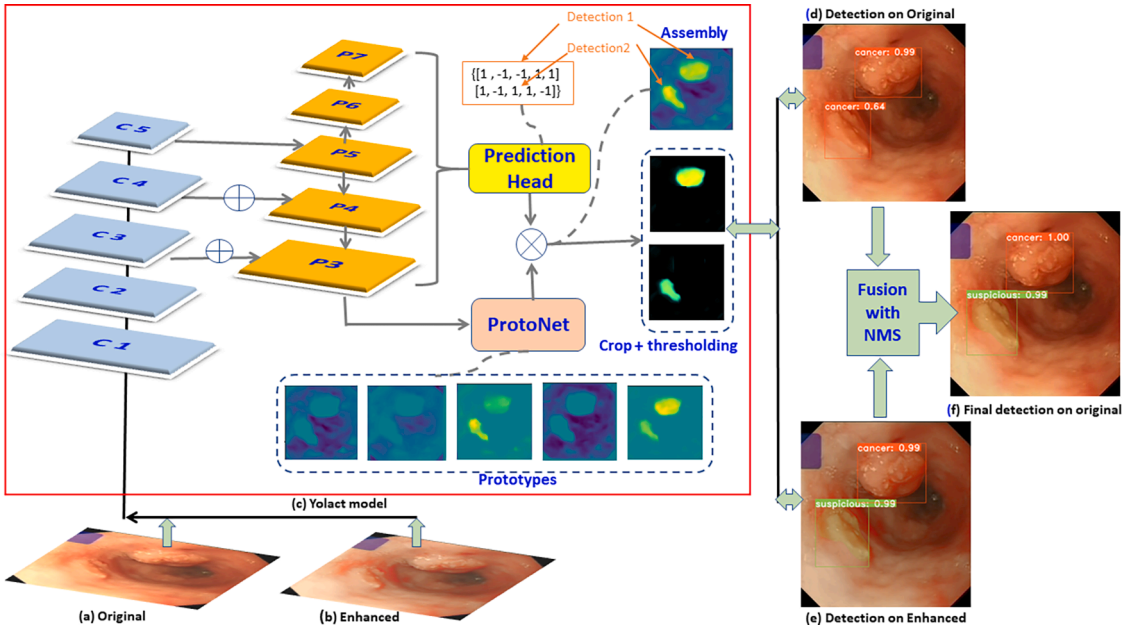


**Fig. 2.** The fused network for detection, delineation and classification of early stage of oesophagus cancer. (a) original image; (b) enhanced image; (c) deep learning system, (d), (e) initial generation of bounding boxes; (f) final fused detection results.

with original images (2(a)). Each image is treated independently. For testing, only the original video is inputted (2(a)). For each frame (2(a)), its enhanced counterpart (2(b)) is generated automatically. Both images are detected one after another (2(d) & 2(e)). Then fusion takes place (by producing combined detected regions in a vector) to construct the final detection superimposed on the original image (2(f)).

For the end-to-end detection system of Yolact (Fig. 2(c)), the basic underline model applies ResNet101 [44] to extract initial feature maps. The object segmentation is accomplished through two parallel subnets (*ProtoNet* and *Prediction Head*), which generates a set of prototype masks

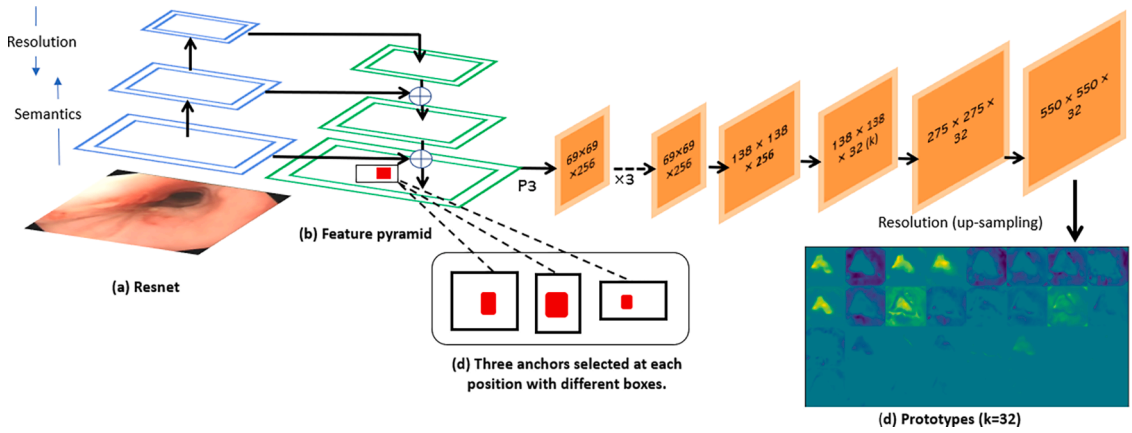


**Fig. 3.** Protonet architecture depicting an image with 550 × 550 pixels with 32 (k) prototypes whereby arrows indicating 3 × 3 *conv* layers. The last layer has *conv* of 1 × 1. (a) The backbone model Resnet for feature extraction with lower resolution indicating higher semantics. (b) Feature pyramid. (c) Three anchors selected at each location with different bounding boxes. (d) Prototypes ($k = 32$) with a full image size (550 × 550).

and predict per-object masque coefficients respectively as explained at Sections 3.1.1 and 3.1.2 respectively.

### 3.1.1. Protonet

Parallel subnet 1, Protonet, in essence, is to generate a dictionary of non-local prototype masks over the entire image as presented in Fig. 3, ProtoNet employs a fully connected network (FCN) accommodating the largest pyramid feature layer (*P*3), to produce a set of image-sized *prototype masks*. These *k* masque prototypes ($k = 32$ in this study, e.g. $[A_1, A_2, \ldots, A_{32}]$) are then applied to deliver predictions for the entire image in relation to classification, segmentation and detection. For *detection* 1 ('cancer') and *detection* 2 ('suspicious') in Fig. 2 with a set of 32 coefficients $[e_1, e_2, \ldots, e_{32}]$ and $[b_1, b_2, \ldots, b_{32}]$ respectively (to be predicted in the Prediction Head in the next section), the two segment marks are calculated as in Eqs. (1) and (2) respectively.

$$mask_1 = e_1A_1 + e_2A_2 + \ldots + e_{32}A_{32} \tag{1}$$

$$mask_2 = b_1A_1 + b_2A_2 + \ldots + b_{32}A_{32} \tag{2}$$

In Fig. 3, the last layer tends to have low resolution but contains strong semantical features. In contrast, the input image has high resolution but with weak semantic features. This protonet operates in a top-down pathway to build prototypes from semantic rich layers.

For an input image with pixels $550(w) \times 550(h)$, the convolution network on 3(a)(b) performs forward pass computing. Since protonet uses input from *P*3 ($69 \times 69$ *pixels*), the deeper backbone layer, the generated masks tend to be more robust. After three more layers with $3 \times 3$ *convolution* (*conv*), the increase in size by up-sampling process will generate $k$ $(= 32)$ full image size prototypes as shown in Fig. 3(d). There are no explicit losses on the prototype masks (more in Section 3.1.3). This *conv* layers from FCN produces $k$ $(= 32)$ masks (Fig. 3(d)) as a matrix $P[w \times h \times k]$. At each location of each feature map, three candidate regions, coined as anchors, with varying sizes and different bounding boxes are selected as potential regions of interest (RoI) for segmentation as elaborated in Section 3.1.2.

### 3.1.2. Prediction head

Parallel subnet 2, entails both predictions of class and bounding box and masque coefficient head for segmentation, which is illustrated in Fig. 4.

Each of five pyramid layers is of square shape with pixel sizes being $69^2$, $35^2$, $18^2$, $9^2$, $5^2$ for $P_3$, $P_4$, $P_5$,$P_6$ and $P_7$ respectively. At each pixel position of each layer, 3 anchors (*A*) are created as candidate RoIs. Hence in total, there will be $19,248$ $(= 3 \times (69^2 + 35^2 + 18^2 + 9^2 + 5^2))$ anchors for each input image. The three anchors have aspect ratio $(AR = w/h)$ of $[1;\ 1/\sqrt{2};\ \sqrt{2}] \times 5$. When AR is 1, the anchor size is $3 \times 3$ where $4.12 \times 2.12$ and $2.12 \times 4.12$ $(4.12 = 3 \times \sqrt{2},\ 2.12 = 3/\sqrt{2})$ are for AR being $1/\sqrt{2}$ and $\sqrt{2}$ respectively. For each anchor, its bounding box is chosen randomly from five pre-defined ones, which have pixels of $(24^2, 48^2, 96^2, 192^2, 384^2)$.

In addition, *Prediction Head* contains three branches, which are c class confidence ($c = 3$ for 'SCC', 'HGD', 'LGD'), 4 bounding box regressors $(=[x_{top-left-corner}, y_{top-left-corner}, x_{bottom-right-corner}, y_{bottom-right-corner}])$, and a vector of masque coefficients, one for each prototype to be processed in parallel. In Fig. 4, the prediction head produces $c$=[0.99, 0.006, 0.004], $box$ = [28, 71, 254, 272], and one 32-masque-coefficient $e$=[1,−1, 1,−1,−1,−1,−1,−1, 1,−1, 1,−1,−1,−1,−1,−1, 0,−1, 0, −1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0] (Fig. 4 (d)).

When these coefficients operate on the prototypes obtained from protonet (Fig. 3)) using Eq. (1), one detection ('Detection 1′) is determined in Fig. 4(e). After crop and threshold (4(f)), together with outcomes from the first 2 branches of *class* and *box*, the final detection is superimposed on the original image (4(g)). Table 2 exemplifies a vector of 32 masque coefficients for the first anchor derived at each pyramid layer *Pi* $(i = 3, 4, 5, 6, 7)$ for the image in Fig. 4. For instance, at $P_3$ layer, there will be 14,283 $(= 69 \times 69 \times 3)$ sets of 32-element vectors, where 3 indicates anchor numbers selected at each feature point.

In summary, all three branches in Fig. 4(c) deliver a vector size of $4 + c + k$ for each anchor (*A*). As a result, for each instance, one or more masks will stem from that instance by linearly combining (plus or minus) the outputs from both prototype and masque coefficient branches (e.g. 'Detection 1′ in Fig. 4(e)), leading to the production of final masks (*M*) (Fig. 4(f)) by a sigmoid nonlinearity as formulated in Eq. (3).

$$M = \sigma(PE^T) \tag{3}$$

where *P* is an $w \times h \times k$ matrix of prototype masks and *E* is a $n \times k$ matrix of masque coefficients for $n$ instances ($n = 2$ ('*detection* 1′ *and* '*detection* 2′) in Fig. 2 and $n = 1$ in Fig. 4) that have passed score thresholding and initial *NMS* as given in Fig. 4(d). In addition, $E^T$ indicates the transpose of *E* matrix.

### 3.1.3. Loss function

The calculation of the loss function is the same as for Yolact [43]. Three loss functions are utilised to train this end-to-end detection model as formulated in Eq. (4), which are classification loss ($\mathscr{L}_{class}$), box regression loss ($\mathscr{L}_{box}$) and masque loss ($\mathscr{L}_{mask}$) where the weights of 1, 1.5, and 1.5 are applied for them respectively to give more weight to classification.

$$\mathscr{L} = \mathscr{L}_{class} + 1.5\,\mathscr{L}_{box} + 1.5\,\mathscr{L}_{mask} \tag{4}$$

In particular,

$$\mathscr{L}_{mask} = BCE(M, M_{gt}) \tag{5}$$

where the binary cross entropy *BCE* is formulated using Eq. (6).

$$BCE(p, y) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \tag{6}$$

where $y$ represents the label and $p$ is the predicted probability of the point being a label for all $N$ points. $M$ and $M_{gt}$ are calculated in Eq. (3).

It should be noted that neither $k$ masque coefficients nor the $k$ prototypes have losses directly occurred on them. Instead, they receive supervision form the final masque loss [43]. For example, if there is a vector $c$ $(1 \times k)$ and a prototype matrix $P$ $(w \times h \times k)$ with ground truth of $gt_{box}$ $(1 \times 4)$ and a binary $gt_{mask}$ $(w \times h \times 1)$, the masque is calculated using Eq. (7).

$$mask = sigmoid(P@c.t()) = sigmoid(P_1c_1 + P_2c_2 + \cdots + P_kc_k) \tag{7}$$

where @ is matrix multiplication and $.t()$ the transpose.

Then the loss is computed as

$$mask\_loss\_tensor = -gt_{mask} * log\,(mask) - - (1 - gt_{mask}) * log(1 - mask) \tag{8}$$

For the purpose of stability, the masque loss if cropped with reference gt box, which is given in Eq. (9).

$$mask\_loss\_crop = crop(mask\_loss\_tensor,\ gt_{box}) \tag{9}$$

Finally, the losses from all RoIs are summed up as

$$\mathscr{L}_{mask} = mask_{loss} = mask\_loss\_crop.sum() \,/\, (gt_{box}.w * gt_{box}.h) \tag{10}$$

As a result, both k prototypes and k masque coefficients receive supervision through Eq. (7).

For back propagation, because derivative of $sigmoid\,(x)$ is $sigmoid(x)(1 - sigmoid(x))$, the derivative of Eq. (7) is formulated in Eq. (11).

$$\nabla mask(P) = mask(1 - mask) * c.t() \tag{11}$$

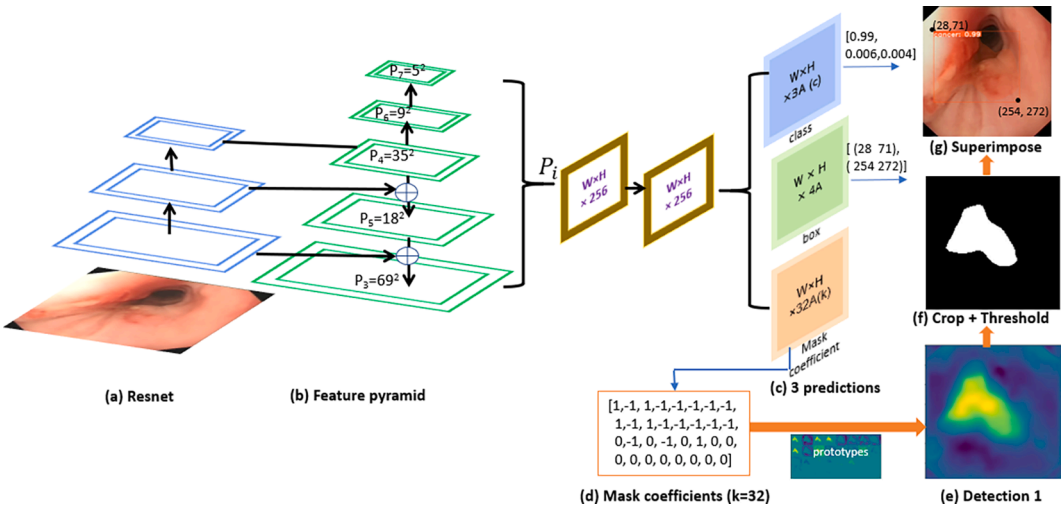


**Fig. 4.** The architecture of prediction head. (a) Backbone model Resnet. (b) Feature pyramid to extract features at each layer where $P_3 = 69^2$, $P_4=35^2$, $P_3=18^2$, $P_6=9^2$, $P_7=5^2$. (c) The network outputs 3 predictions, class, box corner and 32 masque coefficients for each proposed anchor (A).

**Table 2**
An example of 32 masque coefficients (tanh) (for 32 prototypes) for the 5 pyramid layers ($P_3$, $P_4$, $P_5$, $P_6$, $P_7$) for the image shown in Fig. 3. The data listed, are for the first anchor at each layer.

| Pyramid layers | feature size ($w \times h \times 3$) | masque coefficients (tanh) (32, $c_1$, $c_2$, …, $c_{32}$) |
|---|---|---|
| $P_3$ | 69 × 69 × 3 | [0.9175, 0.9317, 0.9872, 0.6649, 0.3953, −0.0503, 0.7729, 0.5816, −0.4834, 0.8479, 0.5081, 0.4353, 0.8370, −0.9635, 0.3952, −0.0851, 0.9728, −0.9942, 0.7079, 0.9088, −0.7926, −0.0494, −0.9997, 0.8667, −0.8194, 0.2543, −0.8955, 0.0538, 0.9622, −0.5045, −0.5384, −0.9494] |
| $P_4$ | 35 × 35 × 3 | [0.2748, 0.9299, 0.7294, 0.9010, 0.9692, 0.9672, 0.8719, −0.7058, 0.5522, 0.6635, −0.1631, 0.2940, 0.7662, −0.9677, −0.8431, 0.3476, 0.8857, −0.8650, 0.4092, 0.2447, 0.1359, −0.4158, −0.9995, 0.2356, −0.5991, 0.2983, −0.8600, −0.8697, 0.3570, 0.8869, 0.0706, −0.9905] |
| $P_5$ | 18 × 18 × 3 | [0.3361, −0.4003, 0.0616, 0.5602, 0.7161, 0.1989, 0.6456, −0.1318, −0.3904, 0.1712, −0.4186, −0.2396, 0.1058, −0.1932, 0.2075, 0.3701, 0.1465, −0.2645, 0.0797, 0.3279, −0.1003, −0.0881, −0.8534, 0.4506, −0.5582, 0.2145, −0.0331, −0.0449, −0.0223, −0.3415, −0.1010, −0.4421] |
| $P_6$ | 9 × 9 × 3 | [0.1814, −0.0925, 0.0594, 0.3306, 0.7110, 0.2590, 0.3865, −0.1222, −0.0477, 0.2004, −0.2092, −0.1247, −0.0115, −0.0665, −0.1179, 0.1331, 0.3571, −0.1150, −0.1220, 0.2321, −0.3143, −0.0296, −0.7765, 0.2023, −0.0271, 0.0126, −0.1410, −0.0694, 0.1867, −0.0024, 0.0726, −0.4259] |
| $P_7$ | 5 × 5 × 3 | [−0.1802, 0.0252, 0.1598, 0.2193, 0.3605, 0.3903, 0.1625, 0.1952, 0.0568, 0.5383, 0.2495, −0.1586, 0.1725, −0.0497, 0.0635, −0.4187, 0.0761, 0.0047, −0.0722, 0.4321, 0.2760, −0.0512, −0.4908, 0.1064, 0.1488, 0.2174, −0.3078, 0.0556, 0.2783, −0.1092, 0.0603, −0.3493] |

the 'loss signal' that for example, prototype 1 ($P_1$), gains, is essentially just weighted by $c_1$ so that the pixels that receive loss are weighted by $mask(1 - mask)$. In other words, if $c_1$ is high and there is a high error, then backpropagation will try to reduce the activations of $P_1$, which is visa versa for negative coefficients.

#### 3.1.4. Implementation

The code is implemented in Python using the PyTorch library under Windows 10 Pro with 64GB RAM and executed using 2 NVidia GeForce GTX 1080Ti GPU cards. The decaying cyclic learning rate (LR) scheme [45] is employed with min and max learning rate $1.3 \times 10^{-4}$ and $1 \times 10^{-3}$ respectively. The cycle length is 50 epochs and at each cycle the max LR decays by a factor of 0.8. A maximum of 500 epochs is trained with early stopping. The batch is set to 4. The ratio between training and validation is set to be 0.9 to 0.1. While this split is initially conducted randomly, manual check follows to ensure that the validation samples contain all three categories (i.e. SCC, HGD and LGD). For testing or evaluation, the independent cohort of subjects are employed, which are not part of training/validation set.

For training, the contrasted images are pre-processed in advance and added to the training set as augmented data (Fig. 2(c)). For testing, upon each input frame, the system generates its contrasted counterpart automatically. After prediction of masks by model in 2(c), the fusion applying NMS with the final detection being presented on the original input frame (Fig. 2(f)).

### 3.2. Generation of contrasted images based on colour appearance model CIECAM

Fig. 5 illustrates the steps of colour contrast enhancement for both WLE and NBI images.

Firstly, the characterisation of the endoscopy camera (5(a)) is performed by recording a palette with 24 standard colours. This palette is also measured using a telespectroradiometer under D65 (average daylight). As such, the correlations between endoscopic camera RGB values (5(d)) and CIE tristimulus values of XYZ (5(f)) is established by a matrix (M) (5(g)). XYZ values are of RGB equivalent calculated from physical colour spectral distributions. Then based on CIE XYZ values, colour appearance model CIECAM is applied to calculate lightness ($J$), colourfulness ($C$) and hue ($H$), for each pixel of an input image (5(h), left most). JCH space represents the colour attributes from human colour perception point of view where $J$ has a range between 0 (no light at all) and 100 (brightest) and $H$ has a circular angle scope with 0 (=360) for

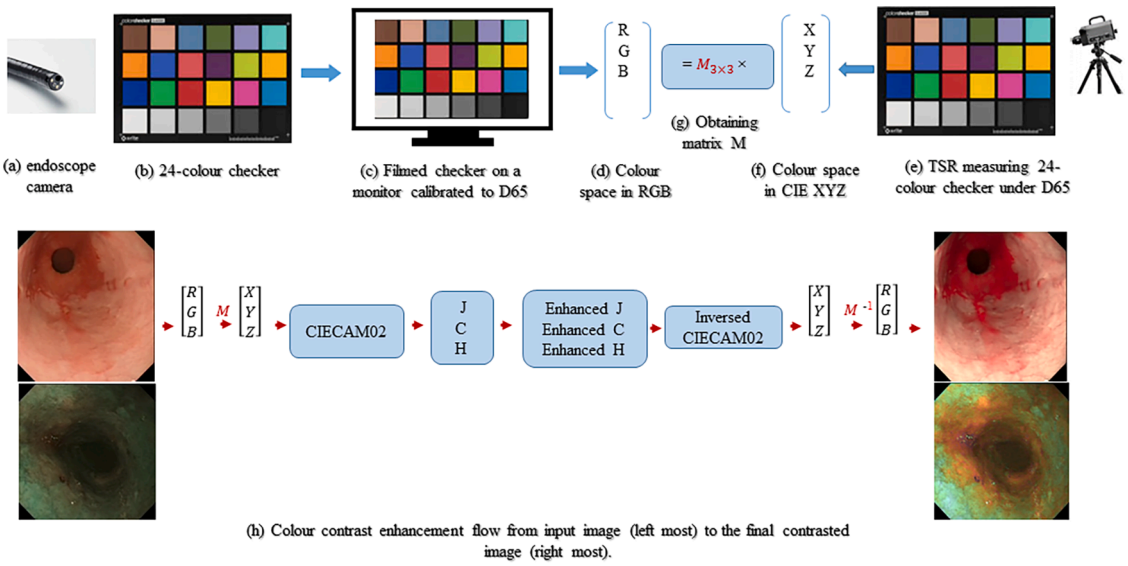


**Fig. 5.** Colour contrast improvement using CIECAM model. (a)–(c) Endoscopy recording of a standard 24-colour checker and measuring the same colour checker under average daylight D65 (e) to obtain the relationship between image RGB values (d) and CIE tristimulus values XYZ (f) by a 3 × 3 matrix (M) for endoscopic cameras (g). (h) Workflow to enhance colour contrast by CIECAM model.

red, 90 for yellow, 180 for green and 270 for blue. Colourfulness $C$ refers to the amount of the concerned hue with 0 indicating no hue (e.g. grey) at all. Although $C$ has no up limit, i.e., as colourful as an object can get, for most endoscopic images with white light or NBI, the maximum $C$ value is around 70 [46].

As demonstrated in Fig. 6 (top row), after a series of measurement taking place for images between diagnosed LGD regions (blue boxes) and their immediate surrounding normal tissues (yellow boxes), the biggest difference appears to occur in colourfulness ($C$) when represented using JCH space. Hence enhancement takes place by simply modifying $C$ values employing Eq. (12).

$$C_{new} = C * C * \beta \tag{12}$$

where $\beta = max(C)/max(C_{new})$, which is to allow small colourfulness being smaller and large being larger, hence to widen the differences in colourfulness, while maintaining the updated colourfulness being consistent with the original value range.

After adjustment of colour attributes, the *JCH* values are converted back to *XYZ* using the inverted matrix $M^{-1}$ (Fig. 5(h)) and then to *RGB* for the final display of enhanced images (Fig. 6 bottom row). Together both original (WLE and NBI) and their enhanced ones are employed to train a deep learning system for detection of cancer (SCC), high grade of dysplasia (HGD) and LG dysplasia (LGD).

The background information used in CIECAM is the averaged epithelium colour measured from normal subjects, for both WLE and NBI under standard viewing environment of D65 as given in Table 3.

In addition, two standard models for measuring colour differences are employed, which are *CIECAM* and $CIEL*a*b*$ (=*CIELAB*) as formulated in Eqs. (13) and (14) respectively.

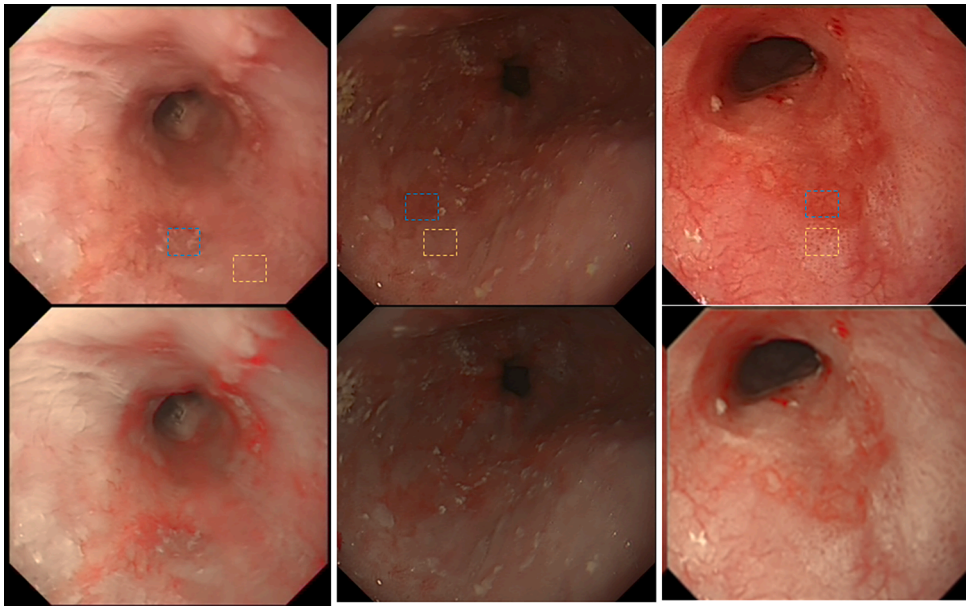

**Fig. 6.** Examples of contrast measurement between known LGD regions (blue dashed box) and their normal tissue surroundings (yellow dashed box). Top row: original white light images. Bottom row: enhance images after applying Eq. (12).

**Table 3**
The averaged RGB and JCH values under D65 for normal subjects.

| | Red (R) | Green (G) | Blue (B) | Lightness (J) | Colourfulness (C) | Hue (H) |
|---|---|---|---|---|---|---|
| WLE | 205 | 136 | 113 | 32 | 16 | 34 |
| NBI | 74 | 83 | 58 | 39 | 9 | 170 |

$$\Delta E_{CAM} = \frac{\sum_n \sqrt{(J_d - J_s)^2 + (C_d - C_s)^2 + (H_d - H_s)^2}}{n} \tag{13}$$

$$\Delta E_{Lab} = \frac{\sum_n \sqrt{(L_d - L_s)^2 + (a_d - a_s)^2 + (b_d - b_s)^2)}}{n} \tag{14}$$

Where subscript $d$ refers to a diseased region (e.g. the blue boxes in Fig. 6 (top row)), $s$ the surrounding region (e.g. the yellow boxes in Fig. 6 (top row) and $n$ the total number of disease-surrounding pairs. The colour attributes, i.e. $L^*$, $a^*$, $b^*$, $J$, $C$, $H$, are the averaged value of each manually selected region (Fig. 6). $H$ in Eq. (13) has been normalised to be within [0,100] from [0,360] to be in the same range with the other two attributes ($J$ and $C$).

### 3.3. Endoscopic datasets

High definition videos including WLE and NBI were collected from patients attending the Translational Gastroenterology Unit at the Oxford University Hospital UK, the Horton General Hospital, Banbury, UK and the Beijing General Hospital, China, using Olympus endoscopes (GIF-H260 or GIF-H290, EVIS Lucera CV260, Olympus Medical Systems, Tokyo, Japan) with recorded videos being in MP4 format, from which still images/frames were extracted. All patients included in this study have given written informed consent to donate biopsies, recording of endoscopic videos and analysis of their clinical data (*REC Ref: 16/YH/0247*).

In total, 389 videos were collected from 389 subjects. Table 4 provides detailed information regarding to the distribution of the collected data sets to the development of concerned AI system. The class (e.g. SCC, HGD, LGD) that each subject was grouped into was based on the worst histological category of that patient as many subjects had multiple lesions with categories of different histology grading, e.g. SCC, HGD and LGD. For training, no normal subject data are included because the background is treated as normal by default to avoid over-fitting.

All videos and images were anonymised by removing all personal information in advance. Two experienced endoscopists with at least 15 years of experience in endoscopic diagnosis and treatment of early oesophageal cancer annotated each image for patients with histologically proven oesophageal squamous neoplasia. The labelling tools were the public software of VGG Image Annotator (VIA)[3] or Amethyst[4] (Zegami, Oxford, UK). Both endoscopists were aware of the histological findings from biopsies taken during the endoscopy.

These images are composed of modalities of WLE and NBI. Corroborated by patients' histology, the surface structure, microvasculature and colour changes of any lesions on images were delineated (i.e. creating masks) and labelled with three classes (suspected dysplasia/low grade dysplasia (LGD), high grade dysplasia (HGD) and cancer (SCC)) using adaptable bounding areas with polygon refinement (bottom row in Fig. 1). The rest non-masque regions were classified as normal (NML), which is a default setting for training as control group.

**Table 4**
Patient numbers (n) that are studied in this work with histological grading of the oesophageal lesions in squamous epithelium.

| Category | SCC | HGD | LGD | Normal | Total |
|---|---|---|---|---|---|
| Training | 29 | 25 | 33 | | 87 |
| Test (independent cohort) | 15 | 17 | 20 | 250 | 302 |
| Total subject | **41** | **42** | **53** | **250** | **389** |

In Table 5, the number of images (video frames) that are for training and testing the developed software system is given. The training takes place based upon still images whereas testing can take either still image or video as an input. For each subject, each video lasts from 10 to 30 min at 30 frames per second, generating 18,000 to 54,000 frames per video. To avoid duplication of same lesions and hence over-fitting, for each subject, frames are selected at different oesophageal locations. Specifically, each video may contain frames of different histology grading, e.g. SCC, HGD and LGD, all of which are included in the experiments. In this collection, cancer images appear to have the smallest number, which may influence the system performance. However, the appearance of a cancer stands out considerably in comparison with that of HGD or LGD, leading to similar accuracy of the test (Table 7) whether cancer data are included or excluded.

In addition, the main purpose of generating contrasted enhanced counterpart is to highlight low grade dysphasia (LGD) to underscore informed information whereas SCC and HGD present more outstanding visual features than LGD. Hence, the evaluation is also provided for detecting LGD only, which sustains a crucial part in identifying patients at risk of developing oesophageal cancer.

It should be noted that the subject number is not the same as sample number. This is because several frames are selected from each subject's video whereas each frame may contain more than one diseased region as illustrated in Fig. 1.

### 3.4. Statistical measures for evaluating

The accuracy of classification and detection are evaluated using common statistical measures of accuracy, recall/sensitivity, specificity whereas segmentation is assessed employing the intersection over union (IoU) as de facto gold standard for evaluating developed computer aided systems. The ground truth is based on the expert annotations in the knowledge of histological findings.

The calculations for system performance are given in Eqs. (15)–(17) [47], where $P$ = *Positive*, $N$ = *Negative*, $TP$ = *True positive*, $FP$ = *False positive*, $TN$ = *True negative*, and $FN$ = *False negative*. Sensitivity or probability of detection assesses the proportion of actual positives that are correctly identified as such. For example, the percentage of 'cancer' regions are correctly labelled as being 'cancer' by the computer system, whereas specificity or true negative rate identifies the proportion of actual negatives (i.e. non-cancer regions) that are correctly labelled as not being cancer. Both specificity and sensitivity are employed for evaluating the performance of classification as well as accuracy.

$$Accuracy = \frac{TP + TN}{P + N} \tag{15}$$

$$sensitivity = recall = \frac{TP}{TP + FN} \tag{16}$$

$$specificity = \frac{TN}{TN + FP} \tag{17}$$

In addition, the overlapping between the boundaries of 2 boxes is quantified using the intersection over union (IoU) as calculated in Eq. (18), which ascertains how much predicted boundary overlaps with the ground truth (the real object boundary).

$$IoU = \frac{area\ of\ overlap}{area\ of\ union} \tag{18}$$

[3] http://www.robots.ox.ac.uk/~vgg/software/via.
[4] https://zegami.com.

**Table 5**
Training and testing data sets in image/frame numbers, where en-WLE=enhanced WLE; en-NBI=enhanced NBI. 3-class=['SCC', 'HGD', 'LGD']. 1-class refers to training with only 'LGD' whereas NML as default.

| Category | SCC | | HGD | | LGD | | Normal | | Total |
|---|---|---|---|---|---|---|---|---|---|
| | Train | Test | Train | Test | Train | Test | Train | Test | |
| WLE for 1 class of LGD | | | | | 339 | 52 | 352 | 60 | **803** |
| WLE + en-WLE for 1 class of LGD | | | | | 678 | 104 | 704 | 126 | **1612** |
| NBI for 1 class of LGD | | | | | 372 | 60 | 352 | 60 | **844** |
| NBI +en-NBI for 1 class of LGD | | | | | 744 | 120 | 704 | 120 | **1688** |
| WLE for 3 classes | 102 | 29 | 112 | 24 | 339 | 52 | | | **658** |
| NBI for 3 classes | 51 | 10 | 156 | 32 | 372 | 80 | | | **733** |
| WLE + en-WLE for 3 classes | 227 | 56 | 262 | 63 | 811 | 140 | | | **1559** |
| NBI + en-NBI for 3 classes | 102 | 20 | 312 | 64 | 744 | 160 | | | **1466** |
| WLE + en-WLE + NBI + en-NBI | 519 | 93 | 1128 | 231 | 2738 | 480 | | | **5189** |

## 4. Results

### 4.1. Visual detection of diseased regions from enhanced regions

Evaluation of proposed enhancement is performed in both visual inspection and training a deep learning system. Visually, an expert clinician annotates 435 images (Table 5) for original WLE images and their corresponding contrasted ones, in comparison with the ground truth (GT) obtained before by different experts. To avoid memory cliché, the evaluation took place over a period of 3 months when each image group (enhanced and original) was annotated interchangeably using the VIA tool. For example, during an annotation day, 50 images in one group and 50 different ones from another group were selected where detection, delineation and labelling were performed. The time spent on each image was accumulated and averaged also provided in Table 6, together with the sensitivity and specificity of classification results for annotating the two groups of images.

While both image groups yield high sensitivity and specificity, enhanced images tend to be more accurately classified with over 98% sensitivity, 3% higher than for the original group and 1.6% higher on specificity. Specifically, the time spent on delineating each colour enhanced frame is 41 s, 14 s (25%) less than the time spent on the original frame (55 s). The averaged overlapping region measured by IoU is 84% for contrasted images, 5% closer to the GT than for the original ones.

Fig. 7 exemplifies a few annotated examples. With regard to lesioned regions, the enhanced group tends to present more detailed boundaries than the original image group, in comparison with GT. The GT regions are also the places where each corresponding biopsy is taken. For example, in the 2nd row, the one blue region on 7(c) was entailed by 2 patches in 7(d). Fig. 7(e) shows the GT.

In particular, most of the images that are overlooked from both original and enhanced groups are of low-grade dysplasia (LGD), highlighting the challenges on detection of precancerous stages, especially within the time constraint during a real time endoscopy. Fig. 8 illustrates this challenge by demonstrating expert's detection in original and enhanced images where the last column provides ground truth.

**Table 6**
Sensitivity (Sen) and specificity (Spe) for the expert labelling image groups of enhanced and original together with the averaged (Avg) time spent on annotating each image and averaged Intersection over Union (IoU) for those correctly classified lesions.

| Image Group | | SCC | HGD | LGD | Avg (%) | Avg time (s) per frame | Avg IoU (%) |
|---|---|---|---|---|---|---|---|
| Enhanced | Sen | 0.9906 | 0.9913 | 0.9590 | 98.03 | 41 | 84 |
| | Spe | 0.9977 | 0.9699 | 0.9970 | 98.82 | | |
| Original | Sen | 0.9626 | 0.9565 | 0.9351 | 95.14 | 55 | 79 |
| | Spe | 0.9931 | 0.9451 | 0.9780 | 97.21 | | |

### 4.2. Evaluation by training a deep learning network

Table 7 supplies the detection results in terms of sensitivity, specificity, and accuracy for the developed deep learning network trained using WLE only, fused with WLE + $WLE_{enhanced}$, NBI only, fused with $NBI$ + $NBI_{enhanced}$, and fused with $WLE + NBI + WLE_{enhanced} + NBI_{enhanced}$. The conventional data colour augmentation approach applies to the training without the fusion by altering RGB values for each input image.

For detection of LGD only with WLE, when fusion is employed, a 7% increase is observed compared to the results without fusion in every measure, i.e. sensitivity, specificity and accuracy. For NBI, around 3% increase is achieved after fusion, which is expected. This is because, in essence, NBI is another form of colour enhancement from WLE by illuminating only blue (415 nm) and green (540 nm) lights. Further contrast enhancement on NBI is only confined to this limited spectral range and may not reveal as much insights as from WLE.

When evaluation for WLE with 3 classes (SCC, HGD, LGD), sensitivity improves by 8% with 4% increase of accuracy when fusion is employed. When both WLE and NBI are applied to train the deep learning system, in average, around 2% increase is observed across all three measures with 3.4% increase in accuracy. For NBI with classification of 3 classes, the increase is marginal (1.1%), implying increasing contrast being more effective for WLE images, the mode that is currently the routine standard for performing endoscopic procedures.

In Table 8, the averaged colour differences are assessed for 100 samples for each of WLE and NBI randomly selected from each LGD region and its surrounding normal mucosa before and after contrast enhancement.

It can be seen that the average difference for enhanced WL is increased significantly at $p < 0.10$ but not for the NBI (with 100 sample pairs) with 14.46 $\Delta E_{L^*a^*b^*}$ and 18.35 $\Delta E_{CAM}$ in comparison with 11.60 $\Delta E_{L^*a^*b^*}$ and 13.12 $\Delta E_{CAM}$ for original images. For $\Delta E_{L^*a^*b}$ measure, human beings cannot perceive any visual difference of 3 or less. Understandably, the enhancement for NBI is not significant ($p>0.1$) as NBI itself is a form of enhancement from WLE by employing the combined lighting at wavelengths of 415*nm* (blue) and 540*nm* (green).

Fig. 9 demonstrates the performance of this developed fused deep learning model on a clip of endoscopic video. The number next to the bounding box refers to the probability of classification, i.e. 'suspicious 0.93′ indicating the delineated region is 93% more likely to be LGD. The bounding boxes are colour-coded with red for 'SCC', blue for 'HGD', and green for 'LGD'.

For the measurement of performance of detection and segmentation, the mean average precisions (mAP) are 67.9% and 59.1% respectively for predicted bounding box and segmentation mark. The Average Precision (AP) is defined as the area under the precision-recall curve. AP is calculated for each class and averaged to get the mAP.

Furthermore, the developed fused system is preliminarily assessed in the Endoscopy unit in Oxford for real-time detection as demonstrated in Fig. 10. The expert endoscopist (10(b)) watches the live endoscopic

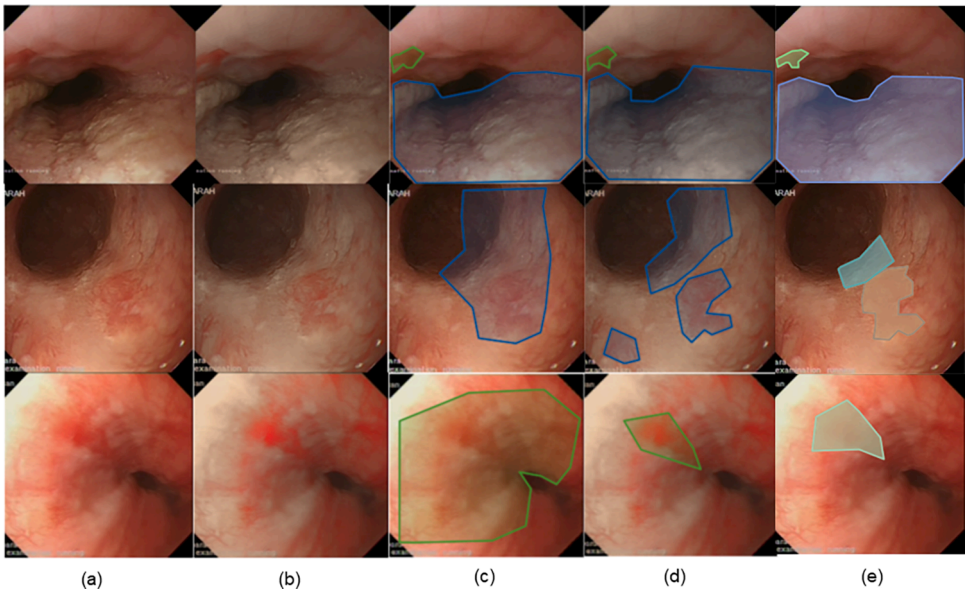


**Fig. 7.** Examples of delineation results from both enhanced and original images by an expert. The ground truth is close to (d). (a) original images; (b) enhanced images; (c) detected region from (a); (d) detected region from (b). Blue=HGD; Green=LGD. (e) ground truth delineated by experts.

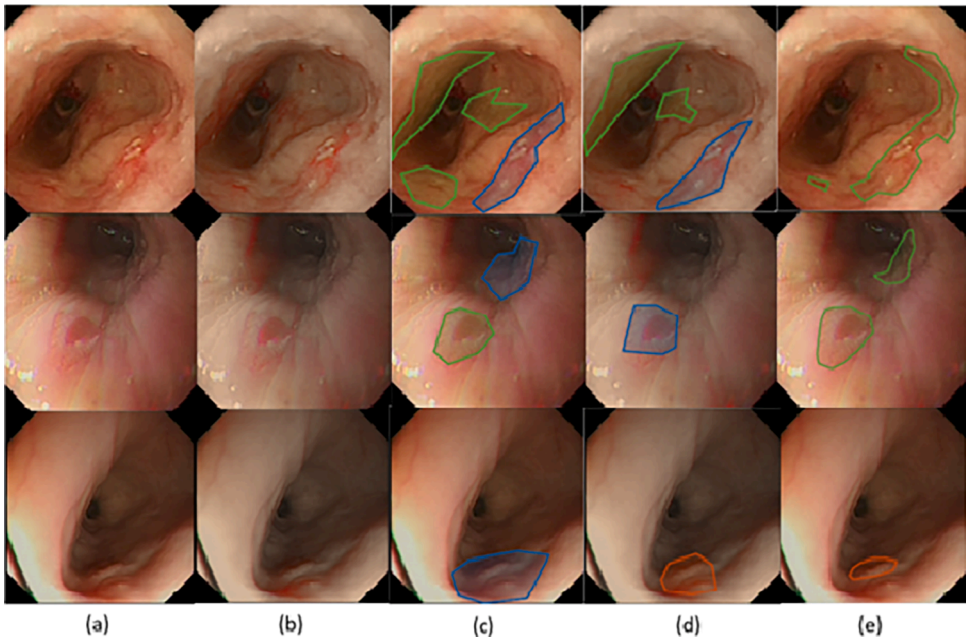


**Fig. 8.** Illustration of sample images with lesions that are missed or wrongly detected from both enhanced and original image groups by experts. (a) original images; (b) enhanced images; (c) detection from original images; (d) detection based on enhanced images; (e) ground truth. Red=cancer, green=LGD, blue=HGD.

video (10(a)) that is firstly transmitted to a laptop in real time (10c)) using a video stream device (StreamCatcher from StarTech.com, Northampton, UK). Then the captured screen is processed and displayed on another monitor (10(d)) with superimposed detection results. The centre red segment on Fig. 10(d) is correctly identified as SCC. This has later been confirmed histologically from the specimen resected during the same endoscopy procedure.

### 4.3. Processing speed

For processing a clip of video, there are two elements to be considered, one is the processing speed and another the continuousness and smoothness of playback of the processed frames. Hence, buffers are employed to process and play back in near parallel fashion to take advantages of computer RAM. With a memory of 64GB in this study, it appears that the setting of 16 buffers delivers the optimal outcome.

To deploy a developed system in a clinical setting, both hardware (e.g., GPU number, computer memory, and monitor size) and software should be considered. A higher number of GPUs, e.g. 2 or 4, will help considerably but will also incur a high cost. Hence, a combination of both, cutting edge hardware systems and optimised algorithms, appears to be the better way forward. Specifically, in Fig. 10, larger monitor sizes as depicted in 10(c)(d) will decrease processing speed.

In this study, Resnet101 is implemented in the system as a backbone for the initial feature extraction of videos (1920 × 1080 pixel/frame), arriving at 33.46 frames per second (fps). For Resnet50 and Darknet53

**Table 7**
The detection results in terms of sensitivity, specificity and accuracy for the deep learning systems trained with and without fusion, including *WLE*, *WLE* + $WLE_{en}$ and *WLE* + *NBI* + $WLE_{en}$ + $NBI_{en}$ images respectively. NML=normal. 1-class refers to training LGD (+ NML) only; 3-class for SCC, HGD, and LGD (+NML). En= enhanced contrast. All the measures are given with standard deviation (± STD).

| Methods | Class | Sensitivity ±STD(%) | Specificity ±STD (%) | Accuracy ±STD (%) |
|---|---|---|---|---|
| WLE for 1-class | LGD | 75.0 ± 2.2 | 88.2 ± 1.7 | 82.7 ± 1.1 |
| | NML | 86.7 ± 3.65 | 80.0 ± 2.4 | 83.3 ± 0.80 |
| **Fusion**: WLE + $WLE_{en}$ for 1 class | LGD | 80.9 ± 5.85 | 84.1 ± 4.1 | 83.0 ± 0.95 |
| | NML | 82.7 ± 4.3 | 96.9 ± 4.35 | 90.6 ± 0.90 |
| NBI for 1-class | LGD | 83.3 ± 1.66 | 90.9 ± 0.5 | 87.3 ± 0.79 |
| | NML | 90.0 ± 3.33 | 85.7 ± 1.19 | 97.7 ± 2.17 |
| **Fusion:** NBI + $NBI_{en}$ for 1-class | LGD | 87.5 ± 2.08 | 93.2 ± 0.8 | 90.3 ± 1.05 |
| | NML | 92.7 ± 0.90 | 88.9 ± 0.59 | 90.6 ± 1.48 |
| WLE for 3-class | SCC | 82.4 ± 3.30 | 95.3 ± 1.07 | 90.5 ± 2.39 |
| | HGD | 74.2 ± 2.91 | 87.3 ± 1.0 | 84.6 ± 1.85 |
| | LGD | 74.5 ± 2.08 | 88.3 ± 1.0 | 83.4 ± 1.32 |
| **Fusion**: WLE + $WLE_{en}$ | SCC | 86.1 ± 2.68 | 93.1 ± 0.94 | 90.3 ± 0.80 |
| | HGD | 87.0 ± 2.47 | 90.2 ± 0.89 | 89.5 ± 0.80 |
| | LGD | 83.0 ± 1.48 | 95.5 ± 1.39 | 90.3 ± 1.02 |
| NBI for 3-class | SCC | 80 ± 0.5 | 99.3 ± 0.2 | 97.4 ± 0.2 |
| | HGD | 87.5 ± 1.6 | 87.5 ± 1.1 | 87.5 ± 0.6 |
| | LGD | 86.7 ± 0.8 | 84 ± 0.85 | 85.5 ± 0.82 |
| **Fusion:** NBI + $NBI_{en}$ for 3-class | SCC | 80 ± 2.5 | 99.5 ± 0.05 | 97.6 ± 0.24 |
| | HGD | 89.1 ± 0.78 | 88.6 ± 1.35 | 88.7 ± 0.75 |
| | LGD | 88.3 ± 2.5 | 85.7 ± 2.47 | 87.1 ± 2.5 |
| WLE + NBI for 3-class | SCC | 87.0 ± 2.5 | 92.6 ± 1.14 | 91.9 ± 1.57 |
| | HGD | 84.3 ± 3.12 | 93.3 ± 4.01 | 86.9 ± 1.70 |
| | LGD | 85.3 ± 2.41 | 90.4 ± 3.41 | 84.2 ± 3.72 |
| **Fusion**: WLE + NBI + $WLE_{en}$ + $NBI_{en}$ | SCC | 90.5 ± 2.23 | 94.9 ± 0.74 | 93.7 ± 0.74 |
| | HGD | 85.0 ± 2.36 | 95.1 ± 1.2 | 88.3 ± 0.88 |
| | LGD | 89.3 ± 2.50 | 93.2 ± 0.87 | 91.4 ± 0.86 |

models, the speeds are 41.43fps and 36.49fps, which are all greater than 24 fps, the minimum frame rate at that human vision cannot perceive motion differences.

In comparison with original Yolact network [43] where 33.5 fps was achieved for processing an image of 550 × 550 pixels with a single Titan Xp graphics processing card (GPU), our work employs 2 GPU cards (Nvidia GeForce GTX 1080Ti) and realises the similar speed with double the size of images (1920 × 1080 pixels). A video clip with classification labels superimposed on the video frames is included in Appendix A. The confidence threshold is set to be 0.3 in processing this video clip.

### 4.4. Explainable aspect of the developed detection model

Analogous to any other software systems, every decision delegated to clinicians calls for clear explanations to ensure its credibility. In this fused system (Fig. 2), this is conducted through an array of prototypes ($k = 32$) with each one presenting the activation status of neural network neurons. Through the linear combination of these prototypes, segmentation masks will be generated. Fig. 11 explains the process where 11(a) shows the images with ground truth and 11(c) the prediction. Fig. 11(b) displays the first twelve prototypes, with each one being the same size as the image itself. Although the number of prototypes can be of any size, it appears that large numbers can make many prototypes redundant for being just blank as exemplified in Fig. 3.

Due to the fact that FCNs are translation invariant, when it comes to the localisation of an object, those translational variances necessitate to be injected back [48] explicitly. In this study, however, with the addition of prototypes, the system learns the way to localise objects via different activations in its prototypes as demonstrated in Fig. 11. Since Resnet101 puts on a rim of padding, the network is able to track the positions of an object and hence is inherently translation variant, the advantage that has been taken in the system. Consequently, the prototypes can also activate on certain 'partitions' of the image as shown with the red dashed line in 11(b). By combining using Eq. (1), e.g. plus or minus, these partition maps, the network can distinguish between different (even overlapping) objects of the same semantic class. Therefore, these prototypes act as an explainable mechanism for the network and fire most strongly on objects that are of interest.

### 4.4. Out-of-sample generalisation

Out-of-sample generalization of disease detection is defined as the ability of an algorithm to achieve similar performance when applied to a completely different institution data or different category dataset [49, 50]. For test of generalisation, the system has been evaluated in a separate data cohort not used for training and development (Tables 4 and 7). In this study, the developed fusion system is also evaluated using images with artefact and with Barrett's oesophagus, which are considered as normal from classification of precancerous stage point of view. Barrett's oesophagus [51] is a premalignant condition with the risk of progression to oesophageal adenocarcinoma. As provided in Table 7 (*WLE-for-1-class* and *NBI-for-1-class*), sensitivity and specificity for these images (classified as 'NML') are 96.8% and 85.2% for WLE image respectively. To evaluate a 3-class system, a clip of Barrett's oesophagus video with 500 frames (1156 × 1912 pixels) is put into a test. For 1-class training, it constitutes normal images, i.e. with artefact and Barrett's, are part of training whereas for 3-class training, the background regions of each delineated images are considered as normal to avoid over-fitting (every lesioned image has a non-lesioned background) with less presence of artefact and Barrett's. Hence, it is expected that classification results are poorer than 1-class system with 98 mis-classified as SCC ($n = 6$), HGD ($n = 2$) and LGD ($n = 90$), leading to an accuracy being 81.6% in comparison with 91.1% for WLE and 90.6% for NBI (Table 7) for 1-class system.

Fig. 12 illustrates an example of processing results (selected at every 200 frames interval from the said video) with 2 frames (arrows) misclassified as LGD.

## 5. Discussion

This work constitutes one of the first to employ fused architecture to improve detection accuracy while overcoming the shortcomings of the existing AI-enhanced decision support systems. Detection of early oesophageal squamous neoplasia remains a challenging task because the surface structure and colour appearance of dysplastic oesophageal mucosa appear inconspicuous to the human eye. Moreover, colour variations in datasets obtained from different centres predominantly render the trained AI-based system only work well with similar datasets when

**Table 8**
The colour differences computed using both $\Delta E_{L^*a^*b^*}$ and $\Delta E_{CAM}$ between each LGD and its surrounding normal mucosa from original WLE and NBI and their enhanced counterparts. All measures are accompanied by a standard deviation (± STD). En=Enhanced.

| | ΔL | Δa* | Δb* | $\Delta E_{L^*a^*b^*}$ | $\Delta L_{CAM}$ | $\Delta C_{CAM}$ | $\Delta H_{CAM}$ | $\Delta E_{CAM}$(%) | p-value (*t*-test) |
|---|---|---|---|---|---|---|---|---|---|
| WLE | 5.76±4.34 | 6.23±4.21 | 7.38±4.80 | 11.60±4.26 | 4.83±2.91 | 9.23±4.44 | 19.30±9.42 | 13.12±9.30 | |
| $WLE_{en}$ | 7.74±4.34 | 8.48±4.22 | 8.44±4.80 | **14.46±3.84** | 6.62±3.32 | 12.29±6.23 | 27.48±14.34 | **18.35±13.48** | **0.071** |
| NBI | 8.17±4.60 | 10.59±1.67 | 9.50±1.93 | 17.52±4.74 | 9.57±6.83 | 3.73±3.04 | 6.25±0.94 | 10.82±6.99 | |
| $NBI_{en}$ | 10.23±6.73 | 16.22±9.76 | 21.56±6.09 | **32.53±8.11** | 12.00±6.95 | 10.06±7.87 | 95.32±18.52 | **33.60±9.02** | **0.162** |

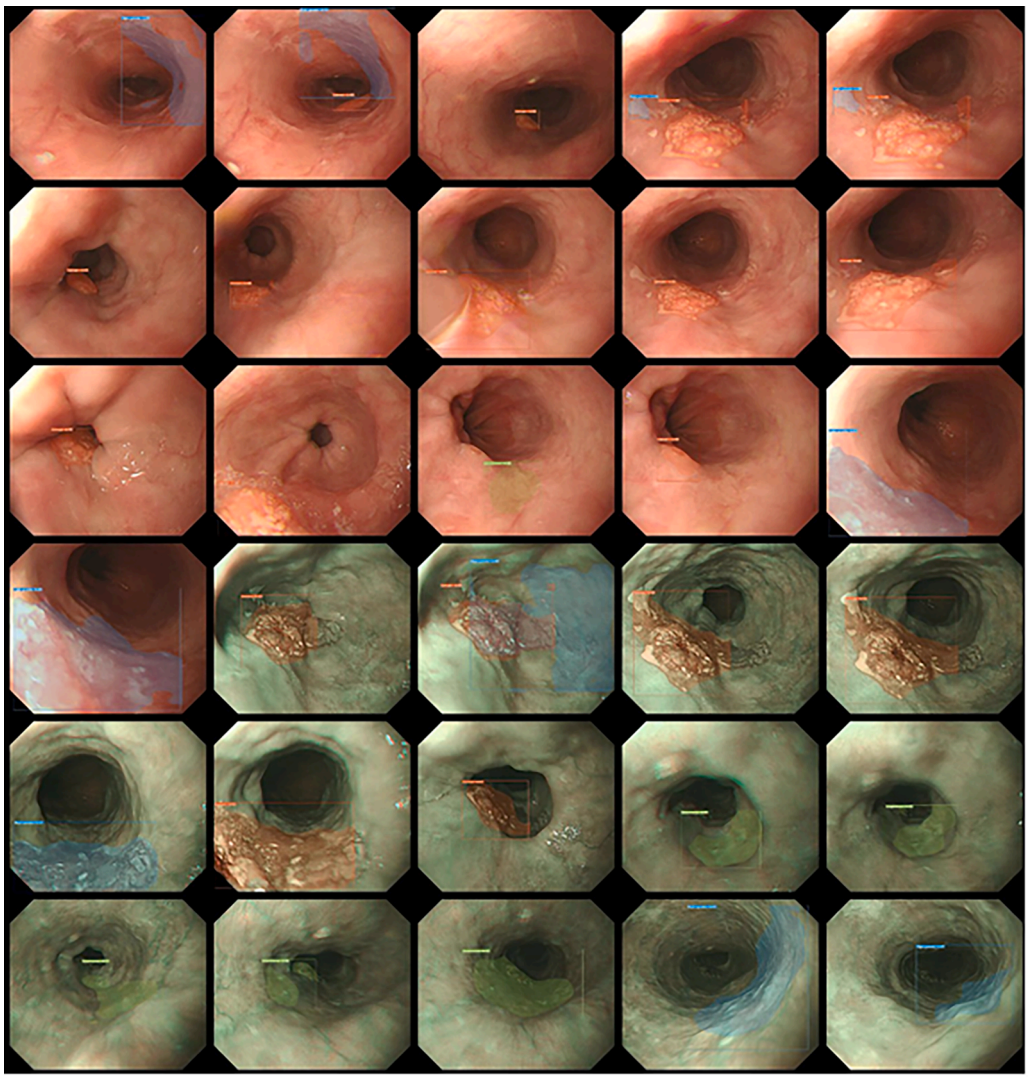

**Fig. 9.** Illustration of detection on a clip of endoscopic video where red='SCC', blue='HGD', green='LGD'.

testing.

Through the revision of colour appearance for contrast enhancement based on human colour vision models, the colour variations between different data sources can be limited to a certain extent, since the contrasted images are yielded under a standard viewing environment of D65 (average daylight) with unified background information (Table 4), aligning contrasted images under similar lighting conditions. In addition, the application of this supervised colour appearance model, CIECAM, to augment data sets, alleviating data shortage appreciably. As a result, dysplastic regions, mainly suspected or LGD, are much more noticeable with colour differences increase from 13.12 to 18.35 in $\Delta E_{CAM}$ for WLE and from 10.82 to 33.60 $\Delta E_{CAM}$ for NBI,. Diagnosing LGD is crucial in identifying patients at risk for developing oesophageal cancer to offer them endoscopic surveillance.

When diagnosing based on enhanced images by an expert clinician, not only is the time (41 s) (Table 6) spent on inspecting each frame 25% less than on the original image, but also the sensitivity, specificity as well as accuracy improved by 3%, 1.5% and 3.5% to being 98%, 98.8% and 98.5% respectively for all three histological grades of squamous neoplasia.

Furthermore, with the addition of these colour contrast-enhanced images to the training and fusion when testing, the accuracy improves from 82.7% to 90.6% for WLE regarding only the LGD class, and to 91.4% when both WLE and NBI images are applied addressing all three histological classes. These results are based on evaluation in an independent cohort of test dataset. Clinically, the most important aspect is to find and identify patients with precancerous alterations of the oesophageal mucosa. Promising in this context is that the sensitivity, specificity and accuracy for detecting LGD are increased from 74.5%, 88.3% and 83.4% to 89.3%, 95.5% and 90.3% respectively when addressing WLE, an improvement by 14%, 7%, and 7% respectively.

For representing colour appearance, CIECAM is an established human vision model simulating human colour perception that is capable to adapt different viewing environments when perceiving an object. Hence, this can leverage the colour differences between different datasets acquired from varying research centres and lead to improved prediction performance of the developed system as all contrasted image frames are created under a standard viewing condition of D65 (average daylight).

In addition, contrasted to the conventional colour augmentation technique, whereby the RGB values are changed linearly at a specific fixed interval, the employment of CIECAM is not only nonlinear, but also true to its original colour. For example, an augmented blue image might not contribute considerably as this colour is not present at current endoscopic procedure.

In comparison with the work conducted by Osawa et al. [22] on colour enhancement based on flexible spectral imaging, where the average increase of $\Delta E_{L^*a^*b^*}$ is 8.4 units from conventional

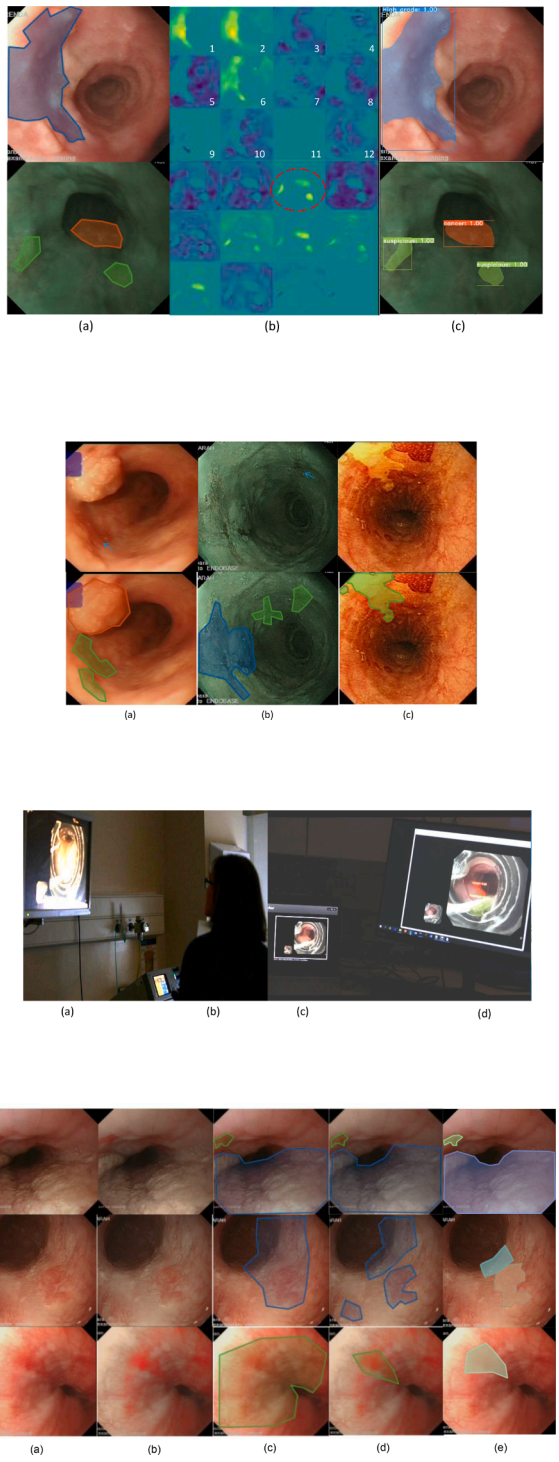


**Fig. 10.** An illustration of an endoscopy procedure in a darkened endoscopy room. (a) The real-time endoscopy video is displayed on a monitor in the endoscopy room. (b) The clinician inspects the endoscopy video while performing endoscopic procedure. (c) Live streaming from the endoscopic system to a laptop which processes the endoscopy data. (d) The detection results are superimposed on the original frame and displayed on a 2nd monitor in real time.

esophagogastroduodenoscopy (EGD) converting WLE to NBI, the contrast in our study has been enhanced by 4.74 and 8.9 units for WLE to WLE and WLE to NBI respectively for their published images, demonstrating the comparable effectiveness of computational technique on image contrast enhancement.

Since CIECAM is a standard model and part of built-in Python library, the conversion from RGB space to JCH space can be performed in a few milli-seconds. For the developed system in this study, to process a clip of videos, the average playing back time after processing stands at 33.46 frames per second (fps) (29 ms per frame (pf)). Significantly, video frames maintain at a high resolution of 1920 × 1080 pixels. At present, for clinical practice and testing, the contrasted images are generated behind the scene whereas only original frames are displayed. Further work will be conducted to show enhanced images as well on the fly, which might require another monitor to depict.

In comparison with recent studies on AI-orientated systems [28–31, 34–39] (Table 1), this developed system exceeds the SOTA results in relation to early detection of squamous cell neoplasia and is probably one of the first tangible real-time detection systems for endoscopic videos for classification of 3 classes, thanks to the inclusion of contrast enhanced images. The deep learning system based on fused contrast enhanced images out-performs with sensitivity, specificity and accuracy being 88.3%, 94.4% and 91.1% respectively for the classification of three histological classes, an increase of 2.8%, 2.3% and 3.4% from the outcomes gained without fusion. When only WLE images are employed for the detection of LGD, contrast enhancement increases the performance by 7.7%, 8.7% and 7.9% respectively. For the calculation of three classes detection, the result of normal (NML) is not included as NML has a much larger proportion of dataset.

There are a number of limitations in this study. Firstly, this fused system is developed using images from only a few centres with limited numbers of training images. Newer endoscope types and processors might provide higher quality images. Secondly, normal oesophagus is set as default in non-annotated areas of training images hence data imbalance might have interfered with the model optimization. However, in the test set we include a large number of controls with normal oesophagus or other diseases (reflux, Barrett's oesophagus). Superiority of the system compared to experts' judgement cannot be demonstrated and this would require prospective clinical studies with targeted biopsies. Thirdly, poor quality images with large amounts of artefacts are excluded in both training and testing sets which might introduce selection bias. Lastly, oesophageal squamous dysplasia and squamous cell carcinoma are the main focus in this study whereas images of Barrett's oesophagus are not analysed. Further studies will investigate this fused system in diagnosing dysplasia in Barrett's oesophagus as well as early oesophageal adenocarcinoma.

A strength of this study is the validation in an external independent cohort of patients from another centre, whereas the controls have included real-world patients with Barrett's oesophagus, reflux oesophagitis, candida oesophagitis and anaemia.

Different from currently published studies, this system can be implemented into a routine clinical setting in an immediate term with little alteration to existing endoscopy equipment (second laptop, a monitor and a video stream catcher required). The standardisation with a defined 24 colour checker facilitates the transfer of the technology to other endoscope video providers. The developed AI-system can guide endoscopists to take targeted biopsies from suspicious lesions which are flagged up on the screen, expectedly leading to minimising the miss rate of early neoplastic lesion during routine endoscopy.

At present, only the attribute of colourfulness of an image is considered, which could potentially limit the application ranges as some samples might present with little alterations in colourfulness but larger changes in other attributes, e.g. lightness. Hence in the future, these attributes will be investigated thoroughly.

## 6. Conclusion

In conclusion, this study introduces a fused real-time multi-modal multi-class endoscopy system, built upon the state-of-the-art artificial intelligence (AI) techniques while assimilating both WLE and NBI imaging modalities and facilitating detection, delineation (segmentation

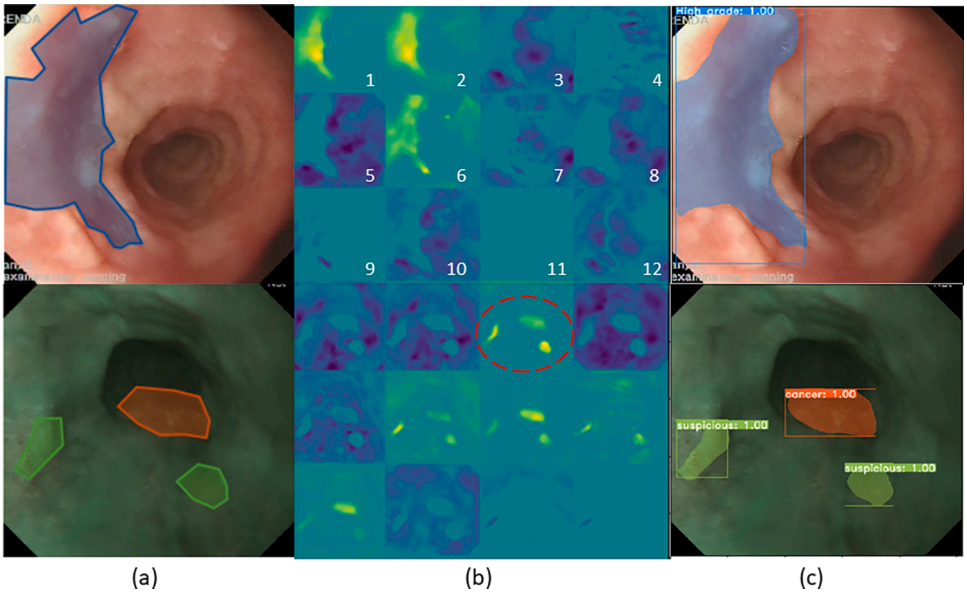


**Fig. 11.** Illustration of explainable nature of trained system in the form of prototypes. (a) Ground truth. (b) Activations of first 12 prototypes. (c) Detection results according to (b). red=’cancer’, blue=’high grade’, green=’suspicious’. Dashed red circle in bottom (b) demonstrates an activated prototype containing *partitions*.

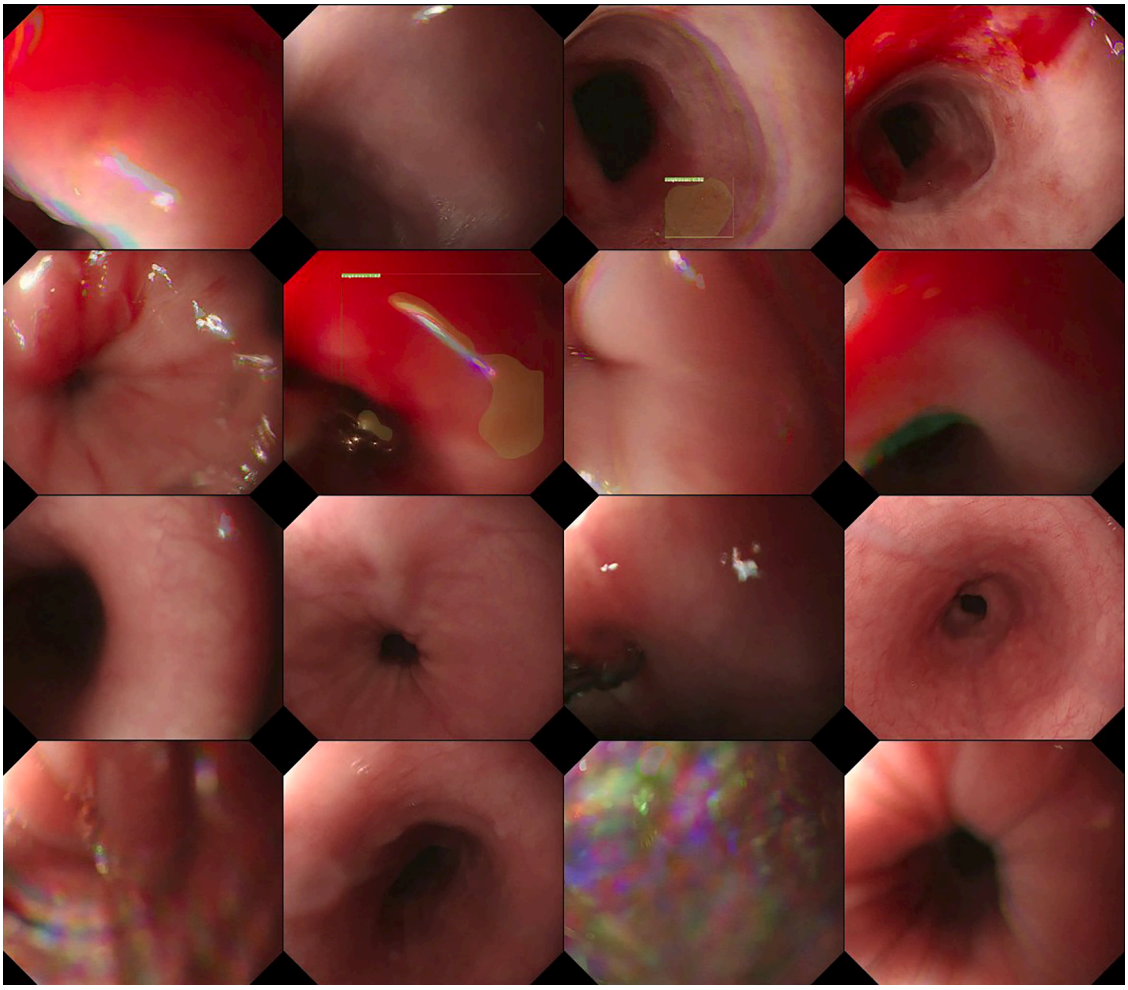

**Fig. 12.** Processing results for a video of Barrett’s oesophagus that is considered as normal in this study. Black arrows indicate the wrong classification.

masks) and characterisation of precancerous (low-grade, high-grade) and cancerous lesions, all at the same time and all in real time.

The developed fused system improves the diagnostic performance and increases the system generalisation. More significantly, colour variations within the datasets obtained from different centres can be leveraged using the contrasted images that are enhanced under standard D65 viewing conditions.

## CRediT authorship contribution statement

**Xiaohong W. Gao:** Methodology, Software. **Stephen Taylor:** Data curation, Software. **Wei Pang:** Methodology. **Rui Hui:** Data curation. **Xin Lu:** Data curation. **Barbara Braden:** Conceptualization, Methodology, Data curation.

## Declaration of Competing Interest

All other authors of this manuscript would like to express no conflict of interest.

## Data availability

Data will be made available on request.

## Acknowledgement

This work was supported by Cancer Research UK [C ref./A 29021] and by the Cancer Research UK (CR-UK) grant number C5255/A18085, through the Cancer Research UK Oxford Centre, and the Oxford Biomedical Research Centre. Their financial support is gratefully acknowledged.

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.inffus.2022.11.023.

## References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J. Clin. 68 (6) (2018) 394–424.
[2] A. Pennathur, M.K. Gibson, B.A. Jobe, J.D. Luketich, Oesophageal carcinoma, Lancet 381 (9864) (2013) 400–412.
[3] M. Arnold, M. Laversanne, L.M. Brown, S.S. Devesa, F. Bray, Predicting the future burden of esophageal cancer by histological subtype: international trends in incidence up to 2030, Am. J. Gastroenterol. 112 (8) (2017) 1247–1255.
[4] M. Arnold, I. Soerjomataram, J. Ferlay, D. Forman, Global incidence of oesophageal cancer by histological subtype in 2012, Gut 64 (3) (2015) 381–387.
[5] R. Siegel, J. Ma, Z. Zou, A. Jemal, Cancer statistics, 2014, CA Cancer J. Clin. 64 (1) (2014) 9–29.
[6] S. Naito, T. Yoshio, A. Ishiyama, et al., Long-term outcomes of esophageal squamous cell carcinoma with invasion depth of pathological T1a-muscularis mucosae and T1b-submucosa by endoscopic resection followed by appropriate additional treatment, Dig. Endosc. 34 (4) (2022) 793–804.
[7] T. Tanaka, S. Matono, N. Mori, K. Shirouzu, H. Fujita, T1 squamous cell carcinoma of the esophagus: long-term outcomes and prognostic factors after esophagectomy, Ann. Surg. Oncol. 21 (3) (2014) 932–938.
[8] G. Chadwick, O. Groene, J. Hoare, R Hardwick, et al., A population-based, retrospective, cohort study of esophageal cancer missed at endoscopy, Endoscopy 6 (7) (2014) 553–560.
[9] T. Chai, X. Jin, S. Li, R. Du, J. Zhang, A tandem trial of HD-NBI versus HD-WL to compare neoplasia miss rates in esophageal squamous cell carcinoma, Hepatogastroenterology 61 (129) (2014) 120–124.
[10] E. de Santiago, N. Hernanz, H. Marcos-Prieto, et al., Rate of missed oesophageal cancer at routine endoscopy and survival outcomes: a multicentric cohort study, United Eur. Gastroenterol. J. 7 (2) (2019) 189–198.
[11] E. Rodriguez de Santiago, N. Hernanz, H.M. Marcos-Prieto, et al., Rate of missed oesophageal cancer at routine endoscopy and survival outcomes: a multicentric cohort study, United Eur. Gastroenterol. J. 7 (2019) 189–198.
[12] W.K. Song, D.G. Adler, J.D. Conway, D.L. Diehl, F.A. Farraye, S.V. Kantsevoy, R. Kwon, P. Mamula, B. Rodriguez, E.J. Shah, W.M. Tierney, Narrow band imaging and multiband imaging, Gastrointest. Endosc. 67 (4) (2008) 581–589.
[13] P.J. Trivedi, B. Braden, Indications, stains and techniques in chromoendoscopy, QJM 106 (2) (2013) 117–131.
[14] Y. Nagami, K. Tominaga, H. Machida, M. Nakatani, N. Kameda, S. Sugimori, et al., Usefulness of non-magnifying narrow-band imaging in screening of early esophageal squamous cell carcinoma: a prospective comparative study using propensity score matching, Am. J. Gastroenterol. 109 (6) (2014) 845–854.
[15] S.P. Oliveira, P.C. Neto, J. Fraga, et al., CAD systems for colorectal cancer from WSI are still not ready for clinical acceptance, Sci. Rep. 11 (2021) 14358.
[16] H. Tizhoosh, L. Pantanowitz, Artificial intelligence and digital pathology: challenges and opportunities, J. Pathol. Inf. 9 (1) (2018) 38.
[17] S. Shahinfar, P. Meek, G. Falzon, How many images do I need? Understanding how sample size per class affects deep learning model performance metrics for balanced designs in autonomous wildlife monitoring, Ecol. Inform. 57 (2020), 101085.
[18] Cho J., Lee K., Shin E., et al., How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?, arXiv:1511.06348, 2016.
[19] C. Zhang, S. Bengio, M. Hardt, B. Becht, O. Vinyals, Understanding Deep Learning (Still) requires rethinking generalization, Commun. ACM 64 (3) (2021) 107–115.
[20] Jimenez-Mesa C., Ramirz J., Sucking J., et al., Deep Learning in current Neuroimaging: a multivariate approach with power and type I error control but arguable generalization ability, arXiv:2103.16685, 2021.
[21] P. Nakkiran, B. Neyshabur, H. Sedghi, The deep bootstrap framework: good online learners are good offline generalizers, in: Proceedings of the ICLR, 2021.
[22] H. Osawa, H. Yamamoto, Y. Miura, H. Ajibe, H. Shinhata, M. Yoshizawa, K. Sunada, S. Toma, K. Satoh, K. Sugano, Diagnosis of depressed-type early gastric cancer using small-caliber endoscopy with flexible spectral imaging color enhancement, Dig. Endosc. 24 (4) (2012) 231–236.
[23] C.L. Tsai, A. Mukundan, C.S. Chung, et al., Hyperspectral imaging combined with artificial intelligence in the early detection of esophageal cancer, Cancers 13 (2021) 4593.
[24] S.E. Martinez-Herrera, et al., Multispectral endoscopy to identify precancerous lesions in gastric mucosa, in: A. Elmoataz, et al. (Eds.), Image and Signal Processing. ICISP 2014. Lecture Notes in Computer Science, Springer, Cham, 2014 vol 85092014.
[25] J. Yoon, J. Joseph, D.J. Waterhouse, et al., A clinically translatable hyperspectral endoscopy (HySE) system for imaging the gastrointestinal tract, Nat. Commun. 10 (2019) 1902.
[26] N. Moroney, M. Fairchild, R. Hunt, C. Li, M. Luo, T. Newman, The CIECAM02 Color appearance model, in: Proceedings of the Tenth Color Imaging Conference IS&T/SID, 2002.
[27] C.J. Li, Z. Li, Z. Wang, et al., Comprehensive color solutions, CAM16, CAT16 and CAM16-UCS, Color Res. Appl. 42 (6) (2017) 703–718.
[28] M. Ohmori, R. Ishihara, K. Aoyama, et al., Endoscopic detection and differentiation of esphageal lesions using a deep neural network, Gastrointest. Endosc. 91 (2) (2020) 301–309, e1.
[29] A.J. de Groof, M.R. Struyvenberg, J. van der Putten, et al., Deep-Learning system detects neoplasia in patients with Barrett's Esophagus with higher accuracy than endoscopists in a multi-step training and validation study with benchmarking, Gastroenterology 158 (4) (2020) 915–929.
[30] Y. Horie, T. Yoshio, K. Aoyama, et al., Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks, Gastrointest. Endosc. 89 (1) (2019) 25–32.
[31] N. Ghatwary, M. Zolgharni, X. Ye, Early esophageal adenocarcinoma detection using deep learning methods, Int. J. Comput. Assist. Radiol. Surg. 14 (2019) 611–621.
[32] Liu W., Anguelov D., Erhan D., Szegedy C., and Reed S., SSD: single shot multibox detector, arXiv:1512.02325, 2015.
[33] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2015, pp. 91–99.
[34] M. Everson, L.C.G.P. Herrera, W. Li, I.M. Luengo, O. Ahmad, M. Banks, et al., Artificial intelligence for the real-time classification of intrapapillary capillary loop patterns in the endoscopic diagnosis of early oesophageal squamous cell carcinoma: a proof-of-concept study, United Eur. Gastroenterol. J. 7 (2) (2019) 297–306.
[35] L. Guo, X. Xiao, C. Wu, et al., Real-time automated diagnosis of precancerous lesion and early esophageal squamous cell carcinoma using a deep learning model (with videos), Gastrointest. Endosc. 91 (2020) 41–51.
[36] H. Mashimo, S.R. Gordon, S. Singh, Advanced endoscopic imaging for detecting and guiding therapy of early neoplasias of the esophagus, Ann. NY Acad. Sci. 1482 (1) (2020) 61–76.
[37] F.L. Dumoulin, F.D. Rodriguez-Monaco, A. Ebigbo, I. Steinbrück, Artificial intelligence in the management of Barrett's esophagus and early esophageal adenocarcinoma, Cancers 14 (2022) 1918.
[38] X.W. Gao, B. Braden, S. Taylor, W. Pang, Towards real-time detection of squamous pre-cancers from oesophageal endoscopic videos, in: Proceedings of the 18th IEEE International Conference on Machine Learning and Applications (ICMLA), 2019.
[39] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 580–587.
[40] K. He, G. Gkioxari, P. Doll´ar, R. Girshick, Mask R-CNN, in: Proceedings of the ICCV, 2017.
[41] Redmon J., Farhadi A., YOLOv3: an incremental improvement, arXiv:1804.02767, 2018.
[42] T.Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal loss for dense object detection, in: Proceedings of the ICCV, 2017.

[43] D. Bolya, C. Zhou, F. Xiao, Y.J. Lee, YOLACT: real-time Instance Segmentation, in: Proceedings of the ICCV, 2019.
[44] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the CVPR, 2016.
[45] Smith, L.N., Cyclical learning rates for training neural networks, arXiv:1506.01186 (2015).
[46] X. Gao, S. Taylor, W. Pang, X. Lu, B. Braden, Early detection of oesophageal cancer through colour contrast enhancement for data augmentation, in: Proceedings of the SPIE Medical Imaging, 2022.
[47] D.G. Altman, J.M. Bland, Diagnostic tests. 1: sensitivity and specificity, BMJ 308 (6943) (1994) 1552.
[48] Y. Li, H. Qi, J. Dai, et al., Fully convolutional instance-aware semantic segmentation, in: Proceedings of the CVPR, 2017.
[49] S. Ali, F. Zhou, B. Braden, et al., An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy, Nat. Sci. Rep. 10 (2020) 2748.
[50] S. Ali, M. Dmitrieva, N. Ghatwary, et al., Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy, Med. Image Anal. 70 (2021), 102002.
[51] J. Fléjou, Barrett's oesophagus: from metaplasia to dysplasia and cancer, Gut 54 (2005) i6–i12.