Hybrid Loss with Network Trimming for Disease Recognition in Gastrointestinal Endoscopy

Qi He[1⋆], Sophia Bano[2], Danail Stoyanov[2], and Siyang Zuo[1]

[1] Key Laboratory of Mechanism Theory and Equipment Design of Ministry of Education, Tianjin University, Tianjin, China
[2] Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS), University College London, London, UK

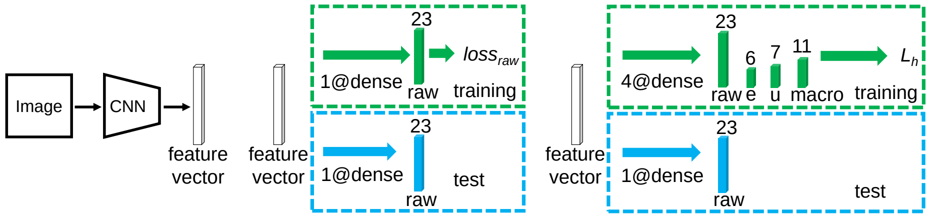
Abstract. *EndoTect Challenge 2020*, which aims at the detection of gastrointestinal diseases and abnormalities, consists of three tasks including Detection, Efficient Detection and Segmentation in endoscopic images. Although pathologies belonging to different classes can be manually separated by experienced experts, however, existing classification models struggle to discriminate them due to low inter-class variability. As a result, the models' convergence deteriorates. To this end, we propose a hybrid loss function to stabilise model training. For the detection and efficient detection tasks, we utilise *ResNet-152* and *MobileNetV3* architectures, respectively, along with the hybrid loss function. For the segmentation task, *Cascade Mask R-CNN* is investigated. In this paper, we report the architecture of our detection and segmentation models and the performance of our methods on *HyperKvasir* and *EndoTect* test dataset.

Keywords: Endoscopy · Object detection · Polyp segmentation · Computer-assisted intervention

1 Introduction

The gastrointestinal endoscopy is a routine examination process via natural cavity for digestive disease detection. It is the most efficient procedure for gastrointestinal disease detection. Although biopsy is the only gold standard for recognising pathology, previous studies on endoscopic imaging reported the potential capability of endoscopy for lesion classification [10, 15]. In these reports, the micro-vascular pattern and micro-surface pattern of the mucosa under the view of endoscopy provided strong evidence for the preliminary diagnosis of gastrointestinal lesion [16]. Well-trained practitioners and experienced endoscopists can detect benign polyps and malignant tumours and tag these lesion with different labels through the micro-anatomical findings visualised by the endoscope. However, these critical clues are unintelligible for a novice practitioner due to their seemly similar appearances. To improve the quality of endoscopy examination, several guidelines have been proposed aiming at quantifying the anatomical

⋆ Corresponding author: Qi He, howard@tju.edu.cn



(a) Backbone **(b)** Single classification head **(c)** Trimmed single classification head

Fig. 1. Proposed hybrid loss with trimming for improving model stability during training. The baseline models are trained using backbone (a) and single classification head (b). $loss_{raw}$ denotes $CE(y_{raw}, \hat{y}_{raw})$. The proposed method with hybrid loss are trained with backbone (a) and multiple classification heads and trimmed to single head during inference (c).

sites to diminish the blind points [1][15]. The recent studies on smart quality control methods based on these guidelines also show their efficiency for endoscopic quality control [7][14]. These computer-assisted lesion detection and anatomical site detection methods showed great potential towards automating the digestive disease diagnosis and endoscopic quality control.

Towards this end, *EndoTect Challenge 2020* (*EndoTect*) called for recognising digestive disease through computer vision methods [8]. The challenge consists of three tasks, namely, detection, efficient detection and segmentation. In this paper, we propose the hybrid loss-based methods utilising *ResNet-152* [6] and *MobileNetV3* [9] for the detection and efficient detection tasks, respectively. The proposed hybrid loss helped in improving the model convergence. For the polyp segmentation task, we use the *Cascade Mask R-CNN* [3] method. Our methods are evaluated on the *HyperKvasir* dataset [2] and the test data of *EndoTect*.

2 Methodology

2.1 Detection and efficient detection

Baseline methods *ResNet-152* [6] and *MobileNetV3-large* [9] are the backbone Convolutional Neural Network (CNN) models that we utilise for the detection and efficient detection tasks, respectively. These models are pre-trained on the *ImageNet* [5] dataset. For fine-tuning, the last fully connected layers are replaced by new dense layers with output units equal to the number of disease classes.

Hybrid loss function We propose a hybrid loss function (L_h) in which the disease labels are rearranged into raw, macro, oesophagus (e) and ulcer (u).

$$L_h = CE(y_{raw}, \hat{y}_{raw}) + CE(y_{macro}, \hat{y}_{macro}) + CE(y_e, \hat{y}_e) + CE(y_u, \hat{y}_u), \quad (1)$$

where CE is the cross-entropy loss. L_h is implemented by adding multiple classification heads after the backbone model as shown in Fig. 1. Corresponding

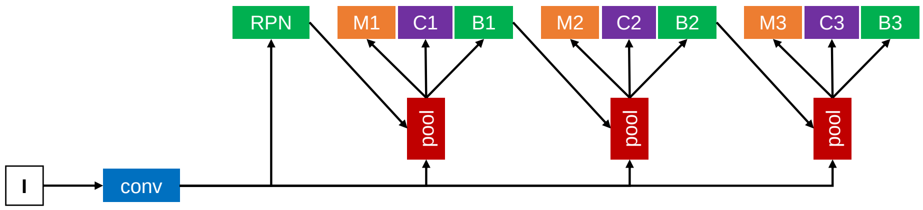


Fig. 2. *Cascade Mask R-CNN*. "I" is input image, "conv" backbone convolution, "RPN" region proposal network, "pool" region-wise feature extraction, "B" bounding box, "C" classification and "M" mask.

labels of y_{raw}, y_{macro}, y_e and y_u for training multiple classification heads are listed in Section 2.3. Our models are trained using the proposed hybrid loss function made up of four cross-entropy loss functions as shown in Fig. 1(b).

Model trimming The multiple classification heads derived from single classification head have a dense layer for the detection task with 23 output units as defined in *EndoTect* and three more dense layers for the extra tasks. This addition of extra units is discussed in more detail in Sec. 4. Though these extra layers improved our model stability during training, they are redundant for inference. Therefore, after training, the multiple classification heads model is trimmed into the single classification head model as shown in Fig. 1(b). This change makes the model lighter and faster during inference.

For brevity, "$\langle model \rangle$" denotes backbone, such as *ResNet-152*, "$\langle model \rangle\, w.$" denotes the model trained with hybrid loss, and "$\langle model \rangle\, w.\, \langle head \rangle$" denotes the classification head, such as *raw*, from model trained with hybrid loss.

2.2 Segmentation model

Our solution is based on *Cascade Mask R-CNN* [3] as shown in Fig. 2, which is implemented using the MMDetection toolbox [4]. The pipeline is formulated as:

$$\begin{aligned} m_t &= M_t(P(x, b_{t-1})), \\ c_t &= C_t(P(x, b_{t-1})), \\ b_t &= B_t(P(x, b_{t-1})). \end{aligned} \tag{2}$$

where x indicate the CNN features of backbone network, $P(.)$ is a pooling operator, e.g., Region of Interest (RoI) Align or RoI pooling, M_t, C_t and B_t denote the mask, class and box head at the t^{th} stage, m_t, c_t and b_t represent the corresponding mask predictions, class predictions and box predictions, respectively. The overall loss function (L_{seg}) takes the form of a multi-task learning:

$$L_{seg} = \sum_{t=1}^{T} (L_{mask}^{t} + L_{bbox}^{t}), \tag{3}$$

$$L_{mask}^{t}(m_t, \hat{m}_t) = BCE(m_t, \hat{m}_t), \tag{4}$$

$$L_{bbox}^{t}(c_t, b_t, \hat{c}_t, \hat{b}_t) = L_{cls}(c_t, \hat{c}_t) + L_{reg}(b_t, \hat{b}_t). \tag{5}$$

Here, L_{mask}^{t} is the loss of mask predictions at stage t, which adopts the binary cross-entropy loss. L_{bbox}^{t} is the loss of the bounding box predictions at stage t, which combines two terms $L_{cls}(c_t, \hat{c}_t)$ and $L_{reg}(b_t, \hat{b}_t)$, respectively for classification and bounding box regression.

2.3 Data augmentation and training details

Data augmentation Training augmentation for detection and efficient detection consists of contrast augmentation, colour shift, brightness augmentation, flipping, perspective transformation and blur. Different from detection, flipping, cutout, colour shift, JPEG compression and affine transform augmentations are applied at random for training the segmentation model.

Labels of hybrid loss The hybrid loss takes label from four categories:

- Raw labels are the original 23 classes provided for *EndoTect*.
- Macro labels consist of 11 classes, namely, 'other', 'bbps-0-1', 'bbps-2-3', 'dyed-lifted-polyps', 'dyed-resection-margins', 'impacted-stool', 'normal-cecum', 'normal-pylorus', 'polyp', 'retroflex-rectum' and 'retroflex-stomach'.
- Oesophagus labels consist of 6 classes, namely, 'other', 'barretts', 'normal-z-line', 'oesophagitis-a', 'oesophagitis-b-d' and 'short-segment-barretts'.
- Ulcer labels consist of 7 classes, namely, 'other', 'ulcerative-colitis-grade-0-1', 'ulcerative-colitis-grade-1-2', 'ulcerative-colitis-grade-2-3', 'ulcerative-colitis-grade-1', 'ulcerative-colitis-grade-2', 'ulcerative-colitis-grade-3'.

Implementation details The detection and efficient detection models are re-implemented with PyTorch [13]. We fine-tuned the models with single GPU for 40 epochs by SGD optimiser with an initial learning rate of 0.003 and momentum of 0.9, and decrease it by 0.1 after 10^{th}, 20^{th} and 30^{th} epochs. The batch sizes for *ResNet-152* and *MobileNetV3* are set to 32 and 128, respectively.

The segmentation model is re-implemented using the MMDetection [4] open-source toolbox based on PyTorch. The model is pre-trained from COCO dataset [12]. Then we fine-tuned it with 2 GPUs for 20 epochs with an initial learning rate of 0.004 and decrease it by 0.1 after 10^{th} and 18^{th} epochs, respectively. The batch size is set to 2 for each GPU. Image data is resized to 1024×1024 pixel resolution for training and inference. For inference, we adjusted the thresholds of the detector. The Non-Maximum Suppression (NMS) threshold of Region Proposal Network (RPN), score threshold of R-CNN, NMS threshold of R-CNN and mask threshold of R-CNN are set to 0.7, 0.5, 0.3 and 0.45, respectively.

Table 1. Average results for detection and efficient detection models

Method	**Dataset**	**Macro Average**			**Micro Average**			
		PREC	REC	F1	PREC	REC	F1	MCC
ResNet-152 raw	HyperKvasir	0.588	0.584	0.584	0.901	0.901	0.901	0.892
ResNet-152 w. raw	HyperKvasir	0.598	0.601	0.596	0.904	0.904	0.904	0.895
ResNet-152 w. raw	EndoTect	0.683	0.646	0.659	0.913	0.913	0.913	0.903
MobileNetV3 raw	HyperKvasir	0.513	0.556	0.504	0.845	0.845	0.845	0.833
MobileNetV3 w. raw	HyperKvasir	0.519	0.557	0.505	0.851	0.851	0.851	0.840
MobileNetV3 w. raw	EndoTect	0.528	0.496	0.503	0.785	0.785	0.785	0.765

3 Results

3.1 Detection and efficient detection

Evaluation metrics consist of precision (PREC), recall (REC) , f1-score (F1) and Matthews correlation coefficient (MCC). We trained and validated *ResNet-152* (*ResNet-152 raw*), *ResNet-152* with hybrid loss (*ResNet-152 w. raw*), *MobileNetV3* (*MobileNetV3 raw*) and *MobileNetV3* with hybrid loss (*MobileNetV3 w. raw*) on *HyperKvasir* dataset following the 2-fold cross validation on the official splits [2]. For *EndoTect*, the models with hybrid loss are trained on *HyperKvasir* and evaluated on the test data provided by *EndoTect*. The models with hybrid loss have an improved performance on *HyperKvasir* than the baseline as shown in Table 1. The *ResNet-152 w. raw* has a superior performance on the images from macro labels than oesophagus labels and ulcer labels, which is demonstrated by the confusion matrix of detection models on *HyperKvasir* as shown in Fig 3.

MobileNetV3 w. is susceptible to the extra black border on the test dataset due to its lighter structure. This is supported by the performance drop of the *MobileNetV3 w. raw* on the test data as shown in Table 1. The test data included dark border regions that were not present in the training data, which made the test data distribution to be slightly different than the training data. These dark borders made the scale of the colour image region on the test data smaller than training data. Though there is some performance drop on it, *MobileNetV3 w. raw* has a great advantage on speed since it has much fewer parameters than *ResNet-152 w. raw*. The speed of *MobileNetV3 w. raw* is evaluated using average time, minimum time, max time, average FPS, minimum FPS and maximum FPS, which are found to be 7.7 ms, 7.6 ms, 22.2 ms, 129.7, 45.0 and 132.0, respectively.

3.2 Polyp segmentation

The segmentation model is evaluated using 2-fold cross validation on *HyperKvasir* dataset. For submission, the model are trained on *HyperKvasir* dataset and evaluated on *EndoTect* test dataset. The evaluation results are shown in Table 2, and the qualitative evaluation is shown in Fig 4. F1-score and Jaccard of 0.879 and 0.822 on the EndoTect test dataset which shows promising performance of our trained model.

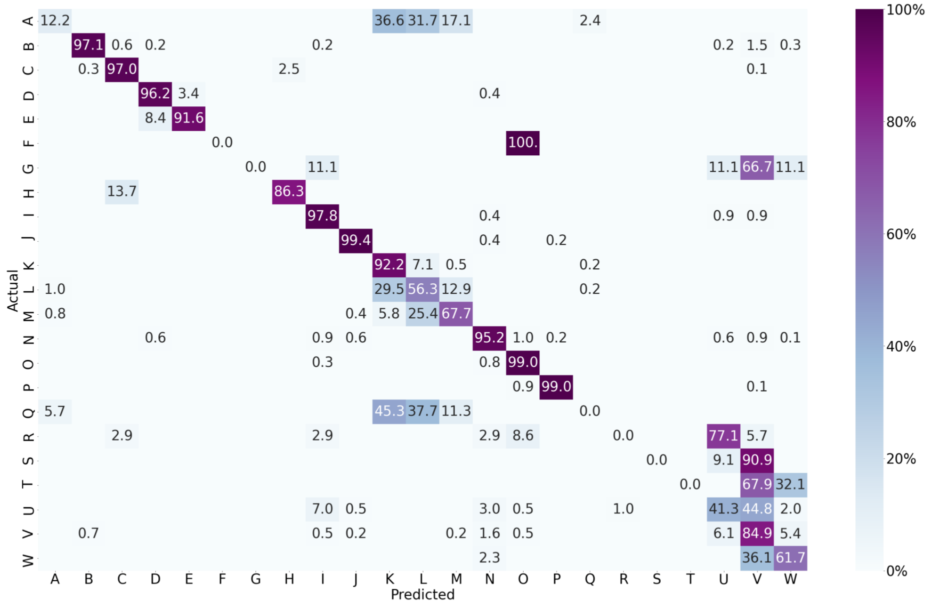
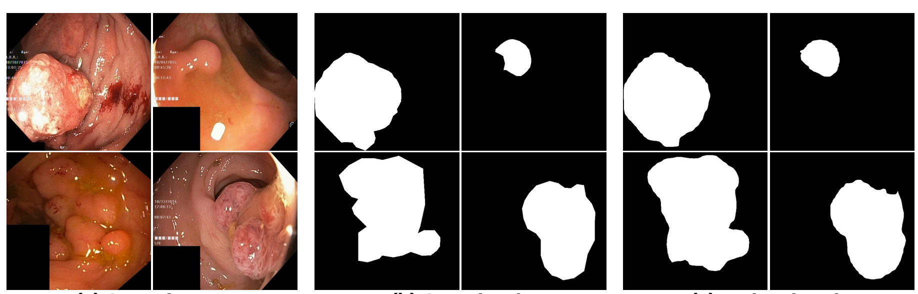


Fig. 3. Confusion matrix of *ResNet-152 w. raw* evaluated on *HyperKvasir*. The labelling of the classes follows [2].

Table 2. Evaluation of segmentation model

Method	Dataset	Jaccard	F1-score	Recall	Precision
Cascade Mask R-CNN	HyperKvasir	0.792	0.850	0.904	0.846
Cascade Mask R-CNN	EndoTect	0.822	0.879	0.882	0.915



(a) Original image **(b)** Ground truth **(c)** Predicted mask

Fig. 4. Qualitative evaluation of the segmentation model

4 Discussion

We proposed a hybrid loss to stabilise convergence of model, which slightly improved the performance of *ResNet-152* and *MobileNetV3* on *HyperKvasir* as shown in Table 1. This change is motivated by an observation, that the CNN model is likely to wrongly classify oesophagus and ulcer images. Such misclassification would last during the whole training process.To narrow the range of misclassification, we filtered these indiscriminate labels through the confusion matrix of *ResNet-152* on *HyperKvasir* and redesign the labels based on the connected component from the confusion matrix. Though focal loss [11] has been demonstrated to achieve a better performance than CE loss in the object detection task, CE loss was found to be experimentally better than the focal loss in this task. Therefore, we designed this hybrid loss (presented in Section 2) using the rearrange labels and CE loss for detection and efficient detection tasks.

Beside the redesigning of labels, we also focused on improving the performance of models via strong image augmentation. After we experimented with various combinations of data augmentation, we found the blur in image augmentation to be detrimental for training segmentation model, because blurring makes it hard to distinguish the features representing boundary and minuscule texture.

5 Conclusion

We addressed the problems of disease detection, efficient disease detection and polyp segmentation for the EndoTect2020 Challenge. We introduced the hybrid loss and model trimming for improving the gastrointestinal disease detection in endoscopic images. The hybrid loss and model trimming is shown to stabilise model training, improve classification of indiscriminate classes and make the model lighter and faster during inference. We utilised Cascade Mask R-CNN with heavy data augmentation for polyp segmentation. We observed that heavy data augmentation helped in better generalising the model for unseen dataset. This was evident from our model superior performance on the EndoTect challenge test dataset compared to the HyperKvasir dataset. The proposed methods are experimentally demonstrated efficient for gastrointestinal image classification and polyp segmentation. In future work, we plan to further improve the multiple classification heads of the hybrid loss for further improving the model performance.

Acknowledgements This work is supported by the Key Technologies Research and Development Program of China (2019YFB1311501), National Natural Science Foundation of China (61773280), Engineering and Physical Sciences Research Council (EP/P027938/1, EP/R004080/1) and Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) at UCL (203145Z/16/Z).

References

1. Beg, S., Ragunath, K., Wyman, A., Banks, M., Trudgill, N., Pritchard, M.D., Riley, S., Anderson, J., Griffiths, H., Bhandari, P.: Quality standards in upper gastrointestinal endoscopy: a position statement of the British Society of Gastroenterology (BSG) and Association of Upper Gastrointestinal Surgeons of Great Britain and Ireland (AUGIS). Gut **66**(11), 1886–1899 (2017)

2. Borgli, H., Thambawita, V., Smedsrud, P.H., Hicks, S., Jha, D., Eskeland, S.L., Randel, K.R., Pogorelov, K., Lux, M., Nguyen, D.T.D., Johansen, D., Griwodz, C., Stensland, H.K., Garcia-Ceja, E., Schmidt, P.T., Hammer, H.L., Riegler, M.A., Halvorsen, P., de Lange, T.: HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Scientific Data **7**(1), 283 (2020)
3. Cai, Z., Vasconcelos, N.: Cascade R-CNN: High Quality Object Detection and Instance Segmentation. arXiv:1906.09756 [cs] (2019)
4. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv:1906.07155 [cs] (2019)
5. Deng, J., Dong, W., Socher, R., Li, L., Kai Li, Li Fei-Fei: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
7. He, Q., Bano, S., Ahmad, O.F., Yang, B., Chen, X., Valdastri, P., Lovat, L.B., Stoyanov, D., Zuo, S.: Deep learning-based anatomical site classification for upper gastrointestinal endoscopy. International Journal of Computer Assisted Radiology and Surgery **15**(7), 1085–1094 (2020)
8. Hicks, S., Jha, D., Thambawita, V., Halvorsen, P., Hammer, H., Riegler, M.: An Overview of the EndoTect Challenge at ICPR 2020. In: Proceedings in the 25th International Conference on Pattern Recognition (ICPR) (2020)
9. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V.: Searching for mobilenetv3. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1314–1324 (2019)
10. Kaise, M., Kato, M., Urashima, M., Arai, Y., Kaneyama, H., Kanzazawa, Y., Yonezawa, J., Yoshida, Y., Yoshimura, N., Yamasaki, T.: Magnifying endoscopy combined with narrow-band imaging for differential diagnosis of superficial depressed gastric lesions. Endoscopy **41**(04), 310–315 (2009)
11. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal Loss for Dense Object Detection. arXiv:1708.02002 [cs] (2018)
12. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft COCO: Common Objects in Context. arXiv:1405.0312 [cs] (2015)
13. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Systems **32**, 8026–8037 (2019)
14. Wu, L., Zhang, J., Zhou, W., An, P., Shen, L., Liu, J., Jiang, X., Huang, X., Mu, G., Wan, X.: Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots during esophagogastroduodenoscopy. Gut **68**(12), 2161–2169 (2019)
15. Yao, K.: The endoscopic diagnosis of early gastric cancer. Annals of Gastroenterology: Quarterly Publication of the Hellenic Society of Gastroenterology **26**(1), 11 (2013)
16. Yao, K.: Zoom gastroscopy: Magnifying endoscopy in the stomach. Springer Science & Business Media (2013)